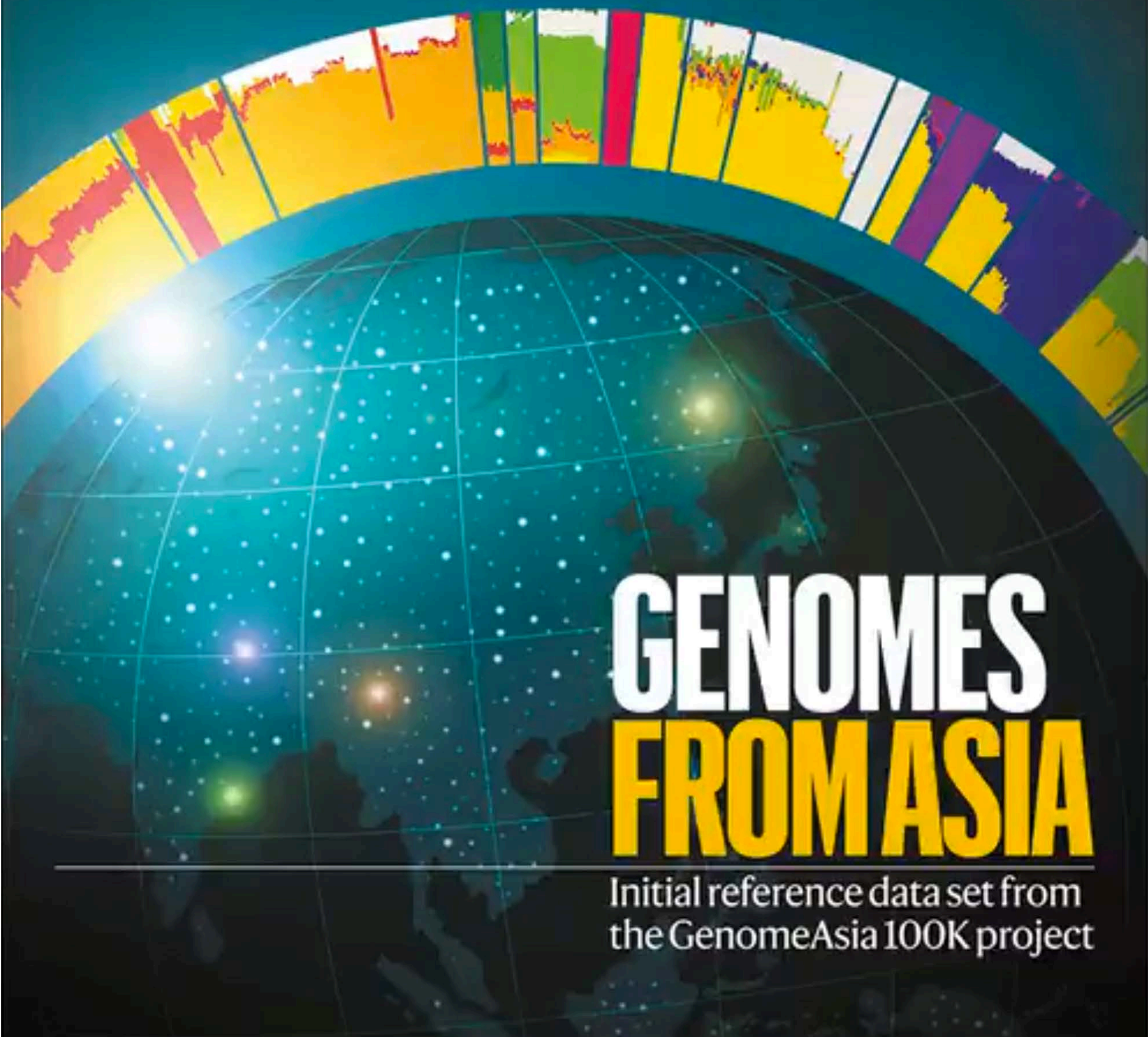


The international journal of science / 5 December 2019



# nature



## GENOMES FROM ASIA

Initial reference data set from  
the GenomeAsia 100K project

### Conference call

Women from some  
minorities get too few  
invitations to talk

### Smoothed layers

Fine-grained metal  
alloy produced by  
3D printing

### Precision changes

Molecular machines  
offer search-and-replace  
genome editing

Vol. 576, No. 7765  
nature.com



## An equitable path to decarbonization

**Madrid climate summit will remain deadlocked unless developed countries accept responsibility for past emissions.**

**T**here is no sign of greenhouse-gas emissions peaking in the next few years.” In an ideal world, such a stark warning – issued by the United Nations Environment Programme (UNEP) – would be enough to persuade delegates attending this week’s climate talks in Madrid to take stronger action against the dangers of climate change. But the two-week meeting is unlikely to yield such results. Negotiators representing the world’s governments are more likely to postpone the hard decisions until next year’s talks in Glasgow, UK, when nations are scheduled to improve on the emissions-reduction pledges they set as part of the 2015 Paris climate agreement.

Negotiators need to solve a number of competing problems that date back to the earliest climate talks in the 1990s and for which there are no straightforward solutions.

First, there must be a step-change in efforts to reduce emissions and keep warming to within 2 °C of pre-industrial temperatures. Here, there is halting progress, although momentum is starting to build towards a global commitment to net-zero emissions by 2050.

Emissions from wealthier nations seem to have stabilized, according to the latest UNEP report. But current pledges to reduce emissions are still projected to result in at least 3 °C of warming, and most developed countries are not even on track to meet those commitments.

More drastic reductions must not, however, neglect the development needs of the poorest communities – those lacking access to sufficient food, water, health care and electric power. Progress here has been scant. As we reported in September, developed nations have failed to fulfil their pledges to provide funding to help poorer countries protect themselves. This is despite the fact that it is their past emissions that are contributing to the extreme climate effects. This funding would also enable poorer nations to continue to industrialize, but use less carbon in the process.

In 2010, developed countries pledged US\$100 billion annually by 2020 towards such help. Some \$9.8 billion was pledged in October at a donors’ conference in Paris, but the United States, which is in the process of withdrawing from the Paris agreement, was notable in its absence.

These are some of the reasons why emissions from developing nations show few signs of tailing off. China has only just caught up with developed states, and its per-capita emissions are now close to those of Japan and the European Union. Its emissions from coal are projected to rise further still.

**“Developed nations have failed to fulfil their pledges to provide funding to help poorer countries.”**

The science shows hard truths. If all countries accept the consensus view of scientists, as most say they do, then by 2030, emissions must be no more than 50% of current levels to keep warming to below 2 °C. That would need more than just net-zero emissions by 2050 – and include a swifter end to coal-fired power and the acceleration of renewable energy and electric-vehicle development. Much more funding would also be required, so that developing countries can both decarbonize and protect vulnerable populations.

As campaigners – and, increasingly, younger generations – urge their national delegates to take real action against climate change, they must also urge their governments to back their pledges with cash for the poorest. The tension between ambition to reduce emissions and the demands of equity must be resolved if international climate talks are to reach agreement.

## Tackle sickle-cell economics

**Most people with the disease will not be able to afford the eye-watering costs of treatment.**

**T**here was a time when Olu Akinyanju felt that no one was listening.

In 1994, the physician founded Sickle Cell Foundation Nigeria, with a mission to provide support for people with sickle-cell disease – a hereditary blood disorder that affects 20 million individual worldwide. The condition is most common in tropical regions of sub-Saharan Africa, but is also found in many other parts of the globe. It can cause strokes, organ failure and harrowing episodes of excruciating pain. Between 50% and 90% of children in sub-Saharan Africa and India with the disease will die before their fifth birthday.

For years, Akinyanju tried and failed to get traction with the World Health Organization (WHO). And leading health policymakers in African countries also had other health and development priorities.

Now the landscape is changing. As we describe in a Feature on page 22, sickle-cell disease is finally catching the attention of funders, governments and pharmaceutical companies. But as they work on innovative ways to tackle the disease, one challenge stands out: how to get treatments to those in need.

Most patients come from communities that have long faced discrimination and economic hardship. They can be stigmatized, and discussions about the condition tend to be rare. That’s partly why, although scientists have known the disease’s root molecular cause for 70 years, research has produced few new treatments.

But in the past decade, more support groups have started to spring up in Nigeria. Internationally, organizations ranging from the WHO to the American Society of Hematology



have made treatments for sickle-cell disease a priority. Newborn-screening programmes have been expanding, and efforts are being made to deploy an old chemotherapy drug called hydroxyurea in Africa to help ease symptoms.

Last week, the US Food and Drug Administration (FDA) approved the first drug, voxelotor, to target the cause of the disease. Made by Global Blood Therapeutics in South San Francisco, California, it reduces the interactions between mutated haemoglobin proteins that lead to the sickled blood cells characteristic of the condition. That came hot on the heels of the FDA approving a drug called crizanlizumab, made by Novartis in Basel, Switzerland, which helps to stop the sickled cells from sticking together.

In October, the US National Institutes of Health (NIH) and the Bill & Melinda Gates Foundation in Seattle, Washington, announced a landmark programme to develop gene-based technologies to treat sickle-cell disease and HIV in Africa. Both will contribute US\$100 million over the next 4 years, and the ambition is to fund treatments into clinical trials within 10 years.

These developments are promising, but they don't address one stark reality. Most people with the disease struggle to access even basic health care, and the new treatments have a hefty price tag.

In 2017, the FDA approved a treatment called Endari, made by Emmaus Medical in Torrance, California. Endari is a formulation of the amino acid glutamine, and costs \$13,000 a year. Unsurprisingly, US physicians are struggling to get insurance companies to foot the bill – meaning that many people are unable to access the treatment.

The first gene therapies for the disease, which involve an elaborate procedure much like a stem-cell transplant (see page 18), are likely to cost upwards of \$1 million per patient. And transplant procedures and hospital stays will push costs higher. The excitement even of voxelotor's landmark approval needs to be tempered by the fact that the treatment costs \$125,000 per year per patient.

This means that advocates such as Akinyanju cannot yet slow down. They have made impressive gains. But alongside the growing sums being invested in research and development, foundations, advocates and patients will continue to need support – especially for the costs of treatments.

Researchers can help – not only through their work, but also by continuing to pressure the government officials, donors and health-care providers with whom they interact to consider the issue of who will foot the bill.

The payment question isn't confined to sickle-cell disease. It bedevils many of the bespoke drugs emerging from biomedical research. What is clear is that the current health-care models won't work: insurance companies balk at the costs, and public systems often can't afford them. An answer will require the combined efforts of biomedical scientists, health-care economists, public-health experts and others.

The NIH and the Gates foundation want a future in which the disease can be treated with a one-time therapy in an outpatient setting – and that is potentially achievable. But companies, funders and governments must find ways to ensure that the costs are not shouldered by communities that have already suffered for too long.

“Foundations, advocates and patients will continue to need support – especially for the costs of treatments.”

## Laying the ghost of Icarus

**Humanity is finally getting up close and personal with Earth's nearest star.**

In some ways, NASA's Parker Solar Probe can trace its ancestry to the tale of Icarus, the character from ancient Greek mythology who took flight by donning wings made from feathers and wax. Ignoring advice from his wise father, Daedalus, Icarus flew too close to the Sun, causing the wax to melt, and plunged to his death.

In the spirit of the Icarus legend, the Parker Solar Probe is one of the most daring space missions ever launched, but there's no metaphorical melting wax. The probe's cutting-edge scientific instruments live behind a carbon-composite heat shield 11 centimetres thick that can withstand temperatures of almost 1,400 °C.

The mission's achievements are thanks in no small measure to the work of teams at the Johns Hopkins University Applied Physics Laboratory in Laurel, Maryland, who built the \$1.5-billion probe and designed its trajectory.

The probe was originally supposed to start its journey by flying past Jupiter – the idea being that Jupiter's gravitational influence would hurl it out of the plane of the planets and over the Sun's poles, from where it would record its measurements. But Yanping Guo, a celestial navigator at the Maryland lab, found a way to send it past Venus instead. This, she reasoned, would keep the probe on a path in the planetary plane and would mean the spacecraft could visit the Sun more often and spend more time close to the star. Since its 2018 launch, the probe has passed close to the Sun 3 times – and it will do so another 21 times in the next 6 years, sending back exclusive data from the Solar System's hottest and most dangerous object.

This week, a News & Views article (D. Verscharen *Nature* <https://doi.org/10.1038/d41586-019-03665-3>; 2019) discusses four papers, published in *Nature*, that report the first of the probe's discoveries, resolving mysteries such as the birthplace of the energetic particles that make up the solar wind, which floods interplanetary space.

Astrophysicist Eugene Parker at the University of Chicago in Illinois proposed the existence of the solar wind more than 60 years ago (E. N. Parker *Phys. Fluids* **1**, 171–187; 1958). At that time, few of his peers accepted that he was on to something. Now, at the age of 92, Parker can justifiably revel in the data from the spacecraft named after him.

The Parker Solar Probe has many more solar flybys ahead of it, taking it progressively closer to the star. The spacecraft has yet to cross a long-anticipated boundary into the Sun's corona, or outer atmosphere; beyond that lies a 'here be dragons' realm that no one has ever seen.

The ghost of Icarus has finally been laid to rest. Much more science is sure to come.



# World view

## To fix research assessment, swap slogans for definitions



By Anna Hatch

**Evaluation reforms will go round in circles without conceptual clarity, warns Anna Hatch.**

**T**he need for clarity extends beyond how we communicate science to how we evaluate it. Who can really define stock phrases such as ‘a significant contribution to research’? Or understand what ‘high impact’ or ‘world-class’ mean?

Seven years ago this month, scientists met in San Francisco, California, to call for an end to the practice of assessing research through the impact factors of the journals in which it is published. They demanded that institutions instead be explicit about their criteria and consider all scholarly outputs – preprints, code, data, peer review, teaching, mentoring and so on. Today, thousands have signed the resulting Declaration on Research Assessment (DORA). But actual change is all too slow.

Two years ago, the DORA steering committee hired me to survey practices in research assessment and promote the best ones. Other efforts have similar goals. These include the Leiden Manifesto and the HuMetricsHSS Initiative.

My view is that most assessment guidelines permit sliding standards: instead of clearly defined terms, they give us feel-good slogans that lack any fixed meaning. Facing the problem will get us much of the way towards a solution.

Broad language increases room for misinterpretation. ‘High impact’ can be code for where research is published. Or it can mean the effect that research has had on its field, or on society locally or globally – often very different things. Yet confusion is the least of the problems. Descriptors such as ‘world-class’ and ‘excellent’ allow assessors to vary comparisons depending on whose work they are assessing. Academia cannot be a meritocracy if standards change depending on whom we are evaluating. Unconscious bias associated with factors such as a researcher’s gender, ethnic origin and social background helps to perpetuate the status quo. It was only with double-blind review of research proposals that women finally got fair access to the Hubble Space Telescope. Research suggests that using words such as ‘excellence’ in the criteria for grants, awards and promotion can contribute to hypercompetition, in part through the ‘Matthew effect’, in which recognition and resources flow mainly to those who have already received them.

Many strategies exist to improve equity in academia, but conceptual clarity is paramount. A study probing the use of ‘outcome’ and ‘impact’ in international-development work concluded that such terms undermine evaluation efforts. It proposed a combination of strategies including the use of meaningful qualifiers, such as the type of result and how it relates to a project’s purpose, and the creation of mutually exclusive definitions for terms such as ‘outcome’,

  
**Most assessment guidelines permit sliding standards.”**

**Anna Hatch** is programme director for the Declaration on Research Assessment in Bethesda, Maryland.  
e-mail: ahatch@sfdora.org

‘impact’ and ‘output’ (B. Belcher & M. Palenberg *Am. J. Eval.* **39**, 478–495; 2018).

Some people say that excellence is easy to identify because ‘you know it when you see it’. But Nobel prizes have been awarded for research that was not immediately recognized as a major breakthrough. And it becomes practically impossible to distinguish shades of excellence when many qualified applicants compete for limited funds.

Being explicit about how specific qualities are valued leads assessors to think critically about whether those qualities are truly being considered. Achieving that conceptual clarity requires discussion with faculty members, staff and students: hours and hours of it. The University Medical Center Utrecht in the Netherlands, for example, held a series of conversations, each involving 20–60 researchers, and then spent another year revising its research assessment policies to recognize societal impacts.

Although DORA curates examples of good practice (see [go.nature.com/2qkcssw](http://go.nature.com/2qkcssw)), most of the best efforts cannot (yet) be found in databases or publications. Often the only way to learn about them is through discussion and networking. It was not until DORA held a meeting with the Howard Hughes Medical Institute in Chevy Chase, Maryland, in October that I learnt the University of California, Irvine, had moved to include collaborative scholarship in evaluations. It took an e-mail exchange to learn of an administrator’s personal efforts to find tools that explicitly credit collaboration.

Frank conversations about what is valued in a particular context, or at a specific institution, are an essential first step in developing concrete recommendations. Although ambiguous terms, for instance ‘world-class’ and ‘significant’, are a hindrance when performing assessments, university administrators have also told me that they rely on flexible language to make room to reward a variety of contributions. So it makes sense that more specific language in review, promotion and tenure guidelines must be able to accommodate varied outputs, outcomes and impacts of scholarly work.

The joint meeting of the American Society for Cell Biology and the European Molecular Biology Organization in Washington DC this month will include a mock faculty-recruitment exercise, involving approaches such as removing applicant names and journal titles from bibliographies. Participants will then discuss which standards to apply to improve objectivity, and how to apply them.

Setting such standards will be tough. It will be tempting to fall back on the misleading simplicity of metrics such as impact factors, or on ambiguous terms that can be agreed to by everyone but applied judiciously by no one. It is too early to know what those standards will be or how much they will vary, but the right discussions are starting to happen. They must continue.



# News in brief

## MALARIA CASES DECREASE WORLDWIDE

The number of malaria infections recorded globally has fallen for the first time in several years, according to the World Health Organization (WHO), which published its annual *World Malaria Report* on 4 December.

Rising numbers of cases in 2016 and 2017 sparked fears that progress had stalled in the global fight against the mosquito-borne disease. But the WHO estimates that there were 228 million reported cases in 2018, a decrease of around 3 million from the previous year.

This drop can be attributed in large part to fewer cases in southeast Asia (see 'Malaria in southeast Asia'). The WHO found that, in the past decade, the most marked decline has been in six countries across the Mekong River basin – Cambodia, China, Laos, Myanmar, Thailand and Vietnam.

From 2010 to 2018, malaria cases dropped by 76% in these countries, and malaria-related deaths fell by 95%. In 2018, Cambodia reported zero malaria-related deaths for the first time in the country's history. India also reported a huge reduction in infections,

with 2.6 million fewer cases in 2018 than in 2017.

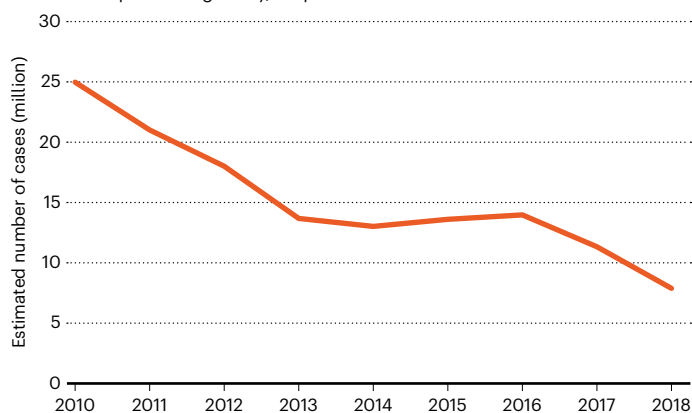
Data on malaria can be inaccurate in countries with poor surveillance systems, warns Arjen Dondorp, deputy director of the Mahidol Oxford Tropical Medicine Research Unit in Bangkok. And even if the number of officially reported deaths is zero, he adds, this doesn't mean that there are no malaria-related casualties. However, "malaria cases are definitely going down" in countries such as Cambodia, he says.

Progress has stalled and even reversed in other parts of the world. Africa, for example, reported an increase of 1 million cases from 2017 to 2018, and the continent accounted for 94% of global cases and deaths from the disease in 2018.

Pedro Alonso, director of the WHO Global Malaria Programme in Geneva, Switzerland, says that, despite the global drop in 2018, malaria cases have stabilized at "unacceptably high numbers" over the past few years. "But this is not a helpless situation," he says, noting that improved efforts to prevent, detect and treat the disease are allowing several countries to successfully eliminate malaria.

## MALARIA IN SOUTHEAST ASIA

The prevalence of malaria has fallen in southeast Asia. Last year, this contributed to an overall drop in cases globally, despite increases in Africa.



## EUROPEAN SPACE BUDGET GETS MASSIVE BOOST

The European Space Agency (ESA) has secured a 45% budget boost. At a meeting in Seville, Spain, on 27–28 November, ministers pledged €12.5 billion (US\$13.8 billion) for 2020–22, compared with the €8.6 billion approved at their 2016 meeting.

ESA's basic-science projects got a 10% hike, the biggest in 25 years. That will allow the agency to bring forward its space-based gravitational-wave mission, the Laser Interferometer Space Antenna (LISA), by two years, from 2034 to 2032, allowing it to observe astrophysical events in tandem with ESA's Athena X-ray telescope, set to launch in 2031.

As part of a new €432-million 'space safety' budget stream, European nations also backed a science and planetary-defence mission. For human and robotic exploration, they earmarked nearly €2 billion, with around €300 million to build modules for NASA's Moon-orbiting Gateway, as well as €150 million for robotic lunar missions.

Meanwhile, Europe's flagship Earth-observation programme, Copernicus (pictured), received €400 million more than the agency had asked for. Other projects that can now press ahead include the design of Europe's first quantum satellite, SAGA, and a project designed to demonstrate ways to remove space debris from orbit.



## Ebola responders killed as violence flares





Armed groups have killed four Ebola responders in the eastern Democratic Republic of the Congo (DRC) and injured seven others in a series of attacks that began late on 27 November, according to the World Health Organization (WHO).

The dead include a vaccination worker, two drivers and a police officer, the agency said. Dozens of aid workers have been evacuated from the areas under siege, and the Ebola response there has mostly halted.

The attacks, in Biakato and Mangina, came after violence in nearby Beni (pictured) prompted the WHO and aid groups to begin evacuating workers from that city last week. “We are heartbroken that people have died in the line of duty,” WHO director-general Tedros Adhanom Ghebreyesus tweeted on 28 November.

One late-night attack targeted the residence of Ebola responders in Biakato. The same night – 27 November – armed groups charged an Ebola-response coordination centre in Mangina.

The violence is poised to drive up the number of new Ebola cases, the WHO said last week. Ebola has killed roughly 2,200 people in the DRC since August 2018.

## CHEMICAL-WEAPONS TREATY BANS NOVICHOKS

The group of nerve agents known as Novichoks are to be added to the Chemical Weapons Convention’s list of controlled substances, in one of the first major changes to the treaty since it was agreed in the 1990s.

The compounds, developed by the Soviet Union during the cold war, came to prominence after they were used in a high-profile assassination attempt on a former Russian military officer, Sergei Skripal, in Salisbury, UK, in March last year.

The Organisation for the Prohibition of Chemical Weapons, which is tasked with enforcing the treaty, announced the decision to explicitly ban Novichoks on 27 November as representatives from the 193 member states met in The Hague in the Netherlands for a periodic review of the convention. The update will come into effect in 180 days.

Novichoks (along with any other nerve agents or deadly chemicals) were already implicitly covered by the convention, which bans the use of any chemical as a weapon. But the specific mention of these compounds in the treaty – and information about their chemical structures – should help to raise global awareness of the ban among chemists.



## CHINESE UNIVERSITIES CLASSSED AS ‘RISKY’ COLLABORATORS

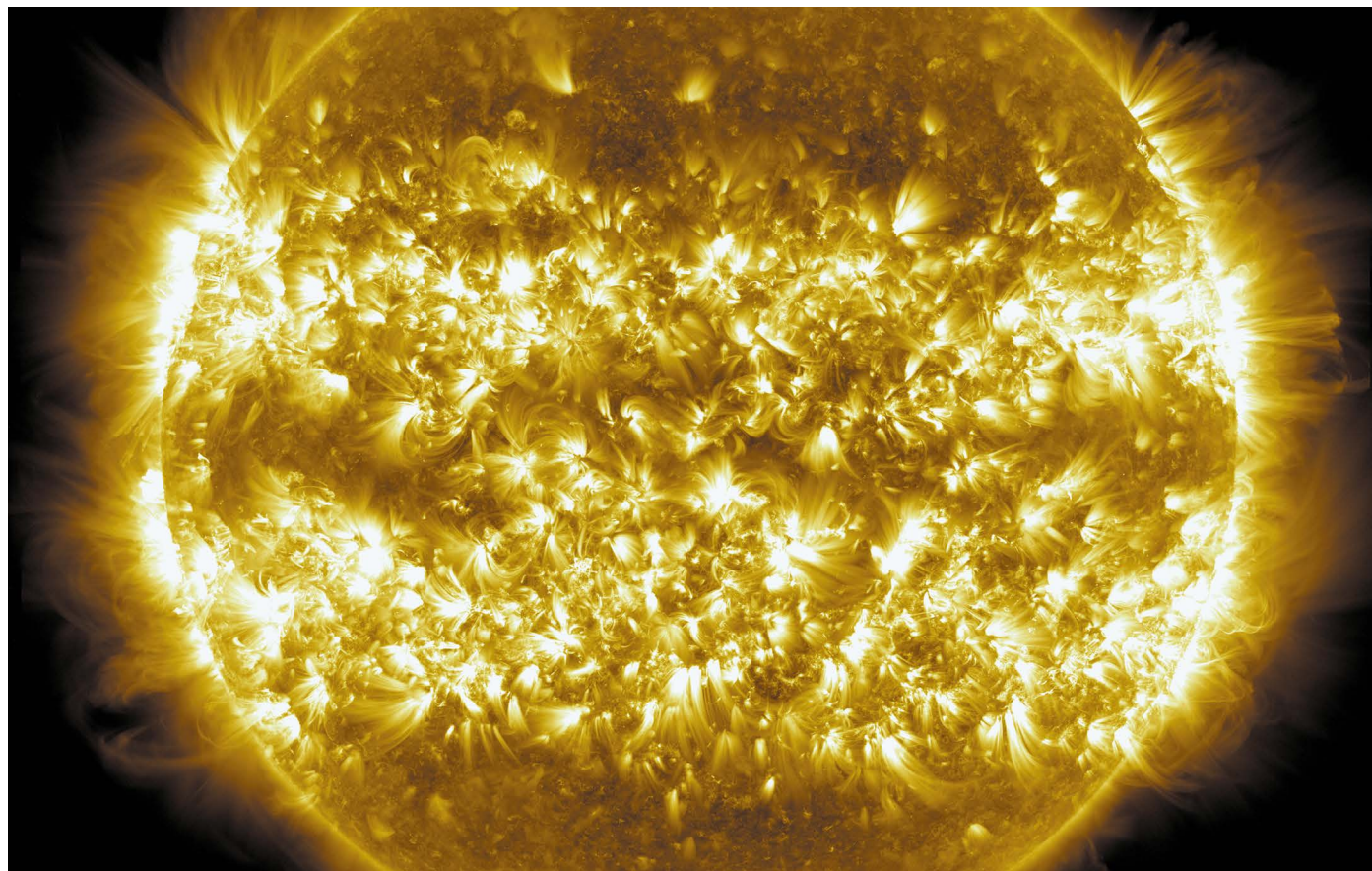
Forty-three Chinese universities are considered ‘very high risk’ or ‘high risk’ collaborators because of their involvement in research for military and defence purposes, according to an Australian think tank. A report published on 25 November by the Australian Strategic Policy Institute in Canberra details how China is using its universities to boost its military prowess.

The institute also launched a database, partly funded by the US State Department, that classifies the level of risk posed by research partnerships with some 160 Chinese universities, security institutions and defence-industry groups. Chinese institutions were included on the basis of their links to defence agencies and the Chinese People’s Liberation Army (PLA) – such as holding security credentials for participating in classified defence or weapons-technology projects, agreements with the PLA or other defence-industry agencies, or records of the institution’s involvement in surveillance.

The analysis comes just weeks after the Australian government released guidelines to help universities reduce the threat of foreign entities, such as the government of China, attempting to leverage activities on campus that are against Australia’s interests.



# News in focus



NASA/SDO/AIA/S. WIESSINGER

Charged particles flow around and away from the Sun. Those that reach Earth can disrupt radio communications.

## SUN-BOMBING CRAFT UNCOVERS SECRETS OF THE SOLAR WIND

Surprise magnetic reversals and a fast rotating wind mark the first findings from NASA's Parker Solar Probe.

By Alexandra Witze

**A** spacecraft buzzing past the Sun has caught the best-ever glimpse of the birthplace of the solar wind – the stream of energized particles that floods outwards from the star.

NASA's Parker Solar Probe spotted strange spikes in the wind, where particles speed up and flip the direction of the wind's magnetic field. The spacecraft also observed the wind rotating around the Sun faster than expected – suggesting that scientists' understanding of how stars' rotation slows down

as they age could be wrong.

The findings, described in four papers published on 4 December in *Nature*<sup>1–4</sup>, could help researchers to better prepare for periods when the solar wind is particularly turbulent and knocks out radio and other communications as it washes over Earth. They are the first discoveries from the Parker Solar Probe, which launched in 2018 and has made three circuits around the Sun so far.

"We're seeing terrific new plasma astrophysics in action, right from the beginning," says Stuart Bale, a plasma physicist at the University of California, Berkeley. "It's been spectacular."

The Parker Solar Probe is gradually drawing closer to the Sun as it loops around the star. The most recent encounter was in September, and the next is expected in January. "We're observing in a regime that we've only speculated about before now," says Sarah Gibson, a solar physicist at the National Center for Atmospheric Research in Boulder, Colorado. The probe is studying the energy that heats the Sun's outer atmosphere, or corona, and accelerates the solar wind.

Although scientists can study the solar wind as it flows over Earth, doing so is like trying to study the origin of a waterfall from halfway

down the cliff over which it pours, says Bale. “If you want to know the source, you have to get up there and get closer – is it coming from one hole in the ground? From a bunch of seams in the rocks? Is there a sprinkler system up there?”

The Parker Solar Probe measured a portion of the solar wind coming from a small hole in the Sun’s corona near the equator<sup>1</sup>. It is the closest look yet at one of the solar wind’s points of origin.

The spacecraft also found that, as the wind streams out into space, parts of it race ahead in high-velocity spikes. “I think of them as rogue waves,” says Justin Kasper, a space scientist at the University of Michigan in Ann Arbor. Within these waves, the speed of the solar wind doubled, and the strong flow temporarily reversed

**“We’re seeing terrific new plasma physics in action, right from the beginning.”**

the wind’s magnetic field<sup>3</sup>. The probe flew through more than 1,000 of these spikes each time it zipped past the Sun, Kasper says. Scientists don’t yet understand what causes them.

Another surprising finding is how quickly the solar wind rotates around the Sun as the star spins. Models suggest that the wind flows in this direction at a speed of a few kilometres per second – but the Parker Solar Probe measured it moving at around 35 to 50 kilometres a second.

The discovery has major implications. Knowing that the wind is rotating at a different speed from expected could help researchers to improve predictions of when a dangerous solar outburst might reach Earth. The finding also suggests that the solar wind is transporting more energy away from the Sun than previously thought, so the star’s rotation might be slowing more rapidly than expected. If so, astronomers might need to revise their ideas about how other stars in the Universe age.

So far, the Parker Solar Probe has studied only a small portion of the Sun at close range. More observations are needed to confirm the unexpectedly fast rotation speed of the solar wind, says Adam Finley, an astronomer at the University of Exeter, UK.

There’s plenty more time for discovery. By the end of its mission in 2025, the probe will have had 24 close encounters with the Sun – getting more than three times closer to the star than it has so far.

1. Bale, S. D. et al. *Nature* <https://doi.org/10.1038/s41586-019-1818-7> (2019).
2. Howard, R. A. et al. *Nature* <https://doi.org/10.1038/s41586-019-1807-x> (2019).
3. Kasper, J. C. et al. *Nature* <https://doi.org/10.1038/s41586-019-1813-z> (2019).
4. McComas, D. J. et al. *Nature* <https://doi.org/10.1038/s41586-019-1811-1> (2019).

# CHINESE MINISTRY INVESTIGATES IMAGES IN TOP ACADEMIC’S PAPERS

Four journals also say they are examining articles co-authored by university president Cao Xuetao.



Cao Xuetao has been a prominent voice for strengthening research integrity in China.

By Andrew Silver

**T**he Chinese education ministry is investigating scientific articles authored by high-profile immunologist and university president Cao Xuetao, following suggestions that dozens of papers contain potentially problematic images. Four journals also say they are examining papers from Cao.

The scrutiny comes after US microbiologist Elisabeth Bik raised concerns three weeks ago on Twitter and the post-publication peer-discussion site PubPeer about images in papers written by Cao and his group.

Cao is the president of Nankai University in Tianjin, and his team has pioneered the development of cancer immunotherapies in China. He says that his group is investigating the papers in question, and he is confident that the issues raised do not affect the papers’ conclusions. Cao has been a prominent voice for strengthening research integrity in China, and gave a speech on the topic at the prestigious Great Hall of the People in Beijing in November.

Bik has flagged up potential problems in about 50 papers co-authored by Cao on PubPeer, and other users, most of them anonymous, have raised similar issues concerning another handful of papers from the group. As *Nature* went to press, images in 63 papers that the team has published in 28 journals since

2003 have been flagged on the site.

In some papers, Bik says, seemingly identical images are labelled as representing different biomedical experiments. In others, features such as patterns of dots that represent biological data seem to be “unexpectedly” duplicated in the same image, she says.

“That would be the equivalent of someone showing you a photo of the night sky, and you would see two Big Dipper constellations in the same photo,” says Bik, who has developed a reputation for spotting and raising potential problems in scientific images and figures.

During a press conference on 22 November, Xu Mei, a spokesperson from China’s Ministry of Education, said the ministry is investigating the articles in question and the “relevant” institutions. Cao is also director of the Institute of Immunology at the Second Military Medical University in Shanghai, also known as the National Key Laboratory of Medical Immunology. Most of the 63 articles list this affiliation.

Representatives from 4 of the 28 journals concerned – *Science*, *Nature Communications*, *Cardiovascular Research* and *Molecular Immunology* – told *Nature* that they had heard about the potentially problematic papers in their journals and were reviewing them.

Bik told *Nature* that she cannot comment on whether the issues she’s flagged up are the result of research misconduct. “It is up to the affiliated institutions to investigate and conclude,” she



says. Although Cao's name is on the papers, often as the corresponding author, it is not clear how closely he was involved in the work.

On 17 November, Cao responded on PubPeer to Bik's comments, saying that his team and collaborators have made it their priority to re-examine the identified manuscripts, raw data and lab records. "We'll work with the relevant journal editorial office(s) immediately if our investigation indicates any risk to the highest degree of accuracy of the published records," he wrote.

He also said he is confident that the conclusions in those papers remain valid and the work reproducible. He apologized for "any oversight on my part" in his role as a mentor, supervisor and lab leader, and added that there is no excuse for a lapse in his supervision or leadership. "I'll use this as an invaluable learning opportunity to do better not only in advancing science, but also in safeguarding the accuracy and integrity of science," he wrote.

Cao did not respond to requests for comment on the issues raised about his team's papers on PubPeer. Nankai University directed *Nature* to Cao's statement on PubPeer.

Individuals, including some who seem to be Cao's co-authors, have responded on PubPeer to some of Bik's queries. In at least one case, a co-author acknowledges that the wrong photograph has been published. In another case, commentators suggest that images flagged as duplicates by Bik were, in fact, pictures of the same cells taken over time, but that the figure's labels were unclear. The explanations given in those cases have been satisfactory, says Bik.

In comments about a few other papers, Bik questions images that the authors have already acknowledged in published errata.

But the authors have not yet responded to questions raised about other papers, in which features such as bars or patterns of dots occur multiple times in the same image, she says.

Several researchers who have not collaborated with Cao or Bik have told *Nature* that the figures she has flagged up seem suspicious. Nicole La Gruta, a molecular biologist at Monash University in Melbourne, Australia, says that, in her opinion: "It is clear from the multiple images that I have seen that these are definitely manipulated."

Wouter Masselink, a postdoctoral molecular biologist at the Vienna BioCenter in Austria, agrees that some of the images require explanation. "I hope the institutions and universities that Cao is associated with launch a formal and independent investigation to find out how and where these artefacts ended up in the published manuscripts," he says.

Bik says she plans to contact the journals that published the papers she has identified. But the comments on Twitter and PubPeer have already caught the attention of some journals. Megan Phelan, a spokesperson for

*Science's* publisher, the American Association for the Advancement of Science in Washington DC, says *Science* is reviewing an article in the journal that Bik flagged up. She added that it's up to institutions to investigate any possible misconduct, which would inform any decisions the journal made.

Elisa De Ranieri, the editor-in-chief of *Nature Communications* in London, says the journal saw posts on Twitter and PubPeer that raised issues over potential image manipulation and will examine any relevant papers as part of their usual research-integrity processes.

Cao received a *Nature* Award in 2015 for excellence in mentoring, and he is co-editor-in-chief of *Cellular & Molecular*

*Immunology*, a journal published by Springer Nature, which also publishes *Nature* (*Nature's* news and comment team is editorially independent of its publisher, and of other Nature-branded journals). A spokesperson for the company says it does not appoint the journal's editorial committee. They said the company is aware that concerns have been raised around some Cao papers but has no further comment.

On 22 November, *Nature Immunology* posted an 'Editor's Note' on two of Cao's papers. One says the authors had flagged up a duplicated image before publication but it was not corrected in time; in the other, the journal says a duplicated image was "inadvertently introduced during the production process".

## UN CLIMATE SUMMIT SET TO TACKLE CARBON MARKETS

Negotiations take place amid uncertain geopolitics and intensifying public pressure.

By Quirin Schiermeier

**F**our years after pledging to limit global warming to no more than 2°C above pre-industrial levels, representatives of nearly 200 countries are meeting to put the finishing touches to the 2015 Paris climate accord.

Discussions at the annual United Nations' climate conference, COP25, are expected to focus on international carbon markets, which have the potential to reduce the overall cost of

global climate-mitigation efforts.

But the talks, which started on 2 December in Madrid and last until 13 December, take place against a backdrop of shifting geopolitics that has created uncertainty over who will lead global efforts to tackle climate change, and of intensifying public pressure on governments to take action.

Despite pledges to curb emissions, atmospheric greenhouse-gas concentrations reached a new peak in 2018, the World Meteorological Organization said last week. A UN



Protesters gather in London as part of the Global Climate Strike in November.

## News in focus

climate report released on 26 November warns that the Paris agreement's 2°C goal might soon be out of reach as emissions continue to rise.

### Unfinished business

At last year's conference, nations agreed on a set of rules for tracking and reporting greenhouse-gas emissions and for reviewing collective progress. However, they failed to establish clear rules for carbon markets through which emissions made in one country can be offset by investing in low-carbon technologies elsewhere. Article 6 of the Paris agreement – which aims to promote voluntary international cooperation between nations – is a central point on the agenda, and offsetting will almost certainly be discussed.

Voluntary offsetting schemes are already in use to make certain goods and services, such as passenger flights, 'carbon neutral'. Many countries, including New Zealand, Sweden and the United Kingdom, rely on offsetting to achieve their emission-reduction goals.

Critics say that offsetting schemes allow rich countries to dodge responsibility for cutting their own emissions. But a well-organized international carbon market with clear, practical rules could save up to US\$250 billion in climate-mitigation costs, says Stefano De Clara, a policy adviser at the International Emissions Trading Association in Brussels. "It would engage businesses in climate action and facilitate the linkage of existing carbon pricing systems," he says. "In the end, everyone could be better off through collaboration."

Analysts have warned that poorly planned offsetting schemes could actually hinder efforts to curb global emissions. Under the Paris agreement, countries must adjust their emission-reduction pledges every five years, in line with the latest scientific evidence about what will be required to stabilize the climate. Without proper rules and bookkeeping, offsetting could simply move emission-reduction efforts around the world, instead of reducing overall emissions, says Gilles Dufrasne, an environmental economist with the Brussels-based international climate-policy watchdog Carbon Market Watch.

Jacob Werksman, a climate-policy adviser at the European Commission, warns that there are some sticking points that negotiators in Madrid might not be able to resolve. For example, some countries expect that excess carbon credits from the expiring 1997 Kyoto Protocol, the previous international climate treaty, will remain eligible for use under the Paris agreement. Such a concession would "severely undermine" the agreement, Werksman says.

This year's talks are also facing intense public scrutiny. The rapidly growing climate-protest movement is shifting the overall conversation on climate change, says Valérie Masson-Delmotte, a co-chair of the Intergovernmental

Panel on Climate Change.

Politics are shifting, too. The United States' official withdrawal from the Paris agreement puts the nation in a strange position for this year's talks. It will remain a member of the UN Framework Convention on Climate Change, an international treaty under which both the Kyoto Protocol and the Paris agreement were negotiated. And US representatives will still attend future COP meetings – including next year's meeting in Glasgow, UK. But unless a future US government revokes the decision to quit the Paris agreement, the country will no longer participate in negotiations concerning

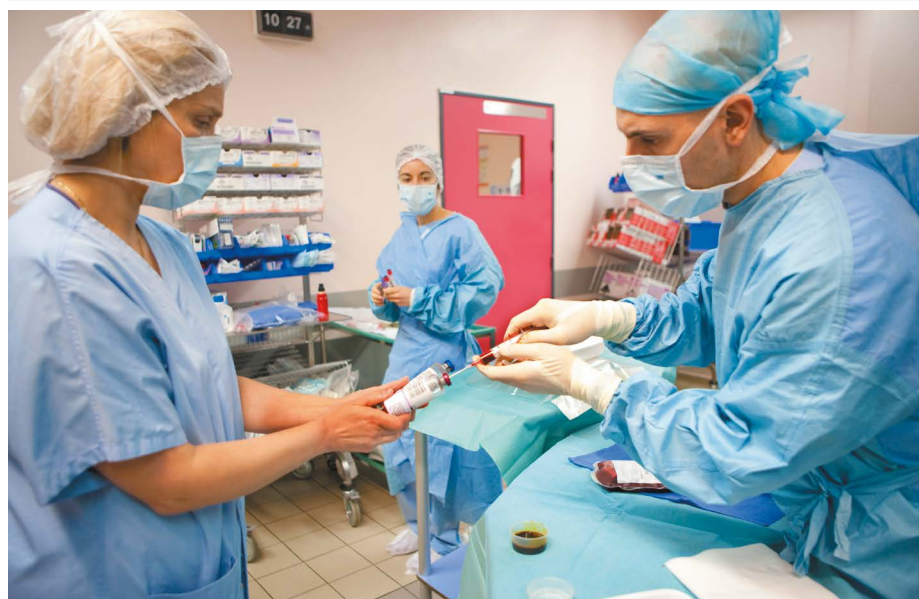
the rules and implementation of the accord.

There is some hope that the European Union will provide new leadership, says Oliver Geden, a policy researcher at the German Institute for International and Security Affairs in Berlin. On 28 November, the European Parliament voted to declare a 'climate and environmental emergency', which will put pressure on EU member states to approve the European Commission's plans to cut emissions by 55% by 2030, and to achieve net-zero emissions by 2050.

"At this time it's up to the EU to demonstrate that the Paris agreement can deliver after all," says Geden. "That's a tough nut to crack."

## TARGETED ATTACKS COULD MAKE BLOOD-STEM-CELL TRANSPLANTS SAFER

Such procedures show promise for genetic and immune disorders, but are currently risky.



Physicians prepare to take a sample of a patient's bone marrow.

By Heidi Ledford

Scientists are experimenting with ways to selectively target the body's blood-making cells for destruction. Early studies in animals and people suggest that the approach could make blood-stem-cell transplants – powerful but dangerous procedures that are used mainly to treat blood cancers – safer, and thereby broaden their use. The studies come as evidence piles up that such transplants can also be used to treat some autoimmune

disorders and genetic diseases.

The work, to be presented at the forthcoming annual meeting of the American Society of Hematology in Orlando, Florida, harnesses an understanding of the proteins made by different types of blood stem cell, the cells in the bone marrow that produce the various cellular components of blood.

Blood-stem-cell transplants work by replacing defective blood-making cells – which can give rise to blood cancer, as well as to genetic and autoimmune diseases – with healthy ones, either from donors or from the patients



themselves. The idea behind the new targeted approaches is to eradicate specific stem cells to make room for transplanted cells without the side effects of existing treatments, which destroy bone marrow cells indiscriminately.

Physicians currently rely on full-body radiation or treatment with toxic, DNA-damaging chemotherapy drugs to kill existing blood stem cells and clear the way for the transplanted cells to repopulate the marrow. That preparation kills not only blood stem cells, but also a host of other cells in the marrow. This can cause infertility, seed cancers that occur later in life, and severely compromise the immune system, leading to lengthy hospital stays.

"It's really prohibitive for patients," says David Scadden, a stem-cell biologist at Harvard University in Cambridge, Massachusetts. "This technology just won't be adopted unless we really change the whole dynamic."

### Stem-cell hotel

One way to think about stem-cell transplants is that the bone marrow is a hotel whose owner wants to evict some guests, says Jens-Peter Volkmer, vice-president of research at Forty Seven, a biotechnology company in Menlo Park, California. Current treatments blow up the whole hotel, he says. "Then everybody's dead, including all of these critical components that you need to protect the patient from infection." The latest approaches allow the owner to tell specific guests to leave – by targeting sets of cells in the bone marrow, rather than killing them all, Volkmer says.

At the haematology meeting, which begins on 7 December, researchers from Forty Seven will present the results of studies that tested a combination of two antibodies in monkeys. One antibody blocks the activity of a molecule called c-Kit, which is found on blood stem cells and is vital to their function; the other inhibits a protein called CD47, which is found on some immune cells. Inhibiting CD47 allows those immune cells to sweep up the stem cells targeted by the c-Kit antibody, making way for new cells.

In the tests, the combination reduced the number of blood stem cells in bone marrow. But the team has not yet demonstrated that the treatment clears out enough old cells to allow transplanted cells to flourish.

Another company, Magenta Therapeutics of Cambridge, Massachusetts, has collaborated with researchers at the US National Institutes of Health to test a different antibody, which binds to c-Kit and then releases a toxin to kill the blood stem cell that produced the protein. Data from studies in mice and one monkey suggest that this can kill off enough stem cells in the bone marrow for transplanted cells to thrive – without destroying other cells such as immune cells.

And a team led by transplant physician Judith Shizuru at Stanford University in California

has tested a similar approach in babies with a genetic disorder that cripples the immune system. The researchers, in a collaboration that includes the firm Amgen of Thousand Oaks, California, used a third antibody that targets c-Kit. The team found that transplanted stem cells, in this case from donors who did not have the disease, successfully took hold in the bone marrow of four out of six of the babies.

### Expanding market

These developments come as the potential market for blood-stem-cell transplants is expanding, says Mani Foroohar, an analyst at SVB Leerink investment bank in Boston, Massachusetts.

Some gene therapies, such as one recently approved by European regulators to treat a genetic immune disorder called ADA-SCID, use a version of the technique. They remove the patient's blood stem cells, then genetically modify them so that they are free of the disorder before infusing them back into the body. Magenta and Forty Seven have entered

into separate collaborations with researchers developing gene therapies to treat blood disorders such as  $\beta$ -thalassaemia and sickle-cell disease (see page 22).

And data are accumulating to show that some people with type 1 diabetes, systemic sclerosis and other autoimmune disorders can enter long-lasting remission if the mature immune cells in their bone marrow are wiped out and replaced with an infusion of their own blood stem cells (E. Snarski *et al. Bone Marrow Transpl.* **51**, 398–402; 2016; K. M. Sullivan *et al. N. Engl. J. Med.* **378**, 35–47; 2018). The procedure is thought to reset the immune system by eradicating cells that are attacking the body's own tissue, says Keith Sullivan, a stem-cell transplant physician at Duke University in Durham, North Carolina.

Sullivan says that the early data from Shizuru and others are intriguing, and that he has begun discussions to collaborate with researchers in the field. "The train is moving now," he says. "The question is, how do we do this in the right way?"

## CARBON DIOXIDE-EATING BACTERIA OFFER HOPE FOR GREEN PRODUCTION

Lab workhorse *E. coli* engineered to make nutrients from greenhouse gas rather than from sugars.

By Ewen Callaway

**E***scherichia coli* is on a diet. Researchers have created a strain of the model bacterium – known as *E. coli* for short – that grows by consuming carbon dioxide instead of sugars or other organic molecules.

The achievement is a milestone, say scientists, because it drastically alters the inner workings of one of biology's most popular model organisms. And, in the future, CO<sub>2</sub>-eating *E. coli* could be used to make organic carbon molecules for biofuels or to produce food.

Products made in this way would have lower emissions than those made using conventional production methods, and could potentially remove the gas from the air. The work was published on 27 November (S. Gleizer *et al. Cell* **179**, 1255–1263; 2019).

"It's like a metabolic heart transplantation," says Tobias Erb, a biochemist and synthetic biologist at the Max Planck Institute for Terrestrial Microbiology in Marburg, Germany, who wasn't involved in the study.

Plants and photosynthetic cyanobacteria – aquatic microbes that produce oxygen – use the energy from light to transform, or fix, CO<sub>2</sub> into the carbon-containing building blocks of life, including DNA, proteins and fats. But these organisms can be hard to genetically modify, which has slowed efforts to turn them into biological factories.

By contrast, *E. coli* is relatively easy to engineer, and its fast growth means that changes

**"After about 200 days, cells capable of using CO<sub>2</sub> as their only carbon source emerged."**

can be quickly tested and tweaked to optimize genetic alterations. But the bacterium prefers to grow on sugars such as glucose – and instead of consuming CO<sub>2</sub>, it emits the gas as waste.

Ron Milo, a systems biologist at the Weizmann Institute of Science in Rehovot, Israel, and his team have spent the past

## News in focus

decade overhauling *E. coli*'s diet. In 2016, they created a strain that consumed CO<sub>2</sub>, but the compound accounted for only a fraction of the organism's carbon intake – the rest came from an organic compound that the bacteria were fed, called pyruvate (N. Antonovsky *et al.* *Cell* **166**, 115–125; 2016).

### Gas diet

In the latest work, Milo and his team used a mix of genetic engineering and laboratory evolution to create a strain of *E. coli* that can get all of its carbon from CO<sub>2</sub>. First, they gave the bacterium genes that encode a pair of enzymes that allow photosynthetic organisms to convert CO<sub>2</sub> into organic carbon.

Plants and cyanobacteria power this conversion with light, but that wasn't feasible for *E. coli*. Instead, Milo's team inserted a gene that lets the bacterium glean energy from an organic molecule called formate.

Even with these additions, the bacterium refused to swap its sugar meals for CO<sub>2</sub>. To further tweak the strain, the researchers cultured successive generations of the modified *E. coli* for a year, giving them only minute quantities of sugar, and CO<sub>2</sub> at concentrations about 250 times those in Earth's atmosphere.

They hoped that the bacteria would evolve mutations to adapt to this new diet. After about 200 days, the first cells capable of using



The model bacterium *Escherichia coli*.

CO<sub>2</sub> as their only carbon source emerged. And after 300 days, these bacteria grew faster in the lab conditions than did those that could not consume CO<sub>2</sub>.

The CO<sub>2</sub>-eating, or autotrophic, *E. coli* strains can still grow on sugar – and would use that source of fuel over CO<sub>2</sub> given the choice,

says Milo. Compared with normal *E. coli*, which can double in number every 20 minutes, the autotrophic *E. coli* are laggards, dividing every 18 hours when grown in an atmosphere that is 10% CO<sub>2</sub>. They are not able to subsist without sugar on atmospheric levels of CO<sub>2</sub> – currently 0.041%.

### A long way to go

Milo and his team hope to make their bacteria grow faster and live on lower levels of CO<sub>2</sub>. They are also trying to understand how the *E. coli* evolved to eat CO<sub>2</sub>: changes in just 11 genes seem to have allowed the switch, and researchers are now working on finding out how.

The work is a “milestone” and shows the power of melding engineering and evolution to improve natural processes, says Cheryl Kerfeld, a bioengineer at Michigan State University in East Lansing.

Researchers have already used *E. coli* to make synthetic versions of useful chemicals such as insulin and human growth hormone. Milo says that his team's work could expand the products the bacteria can make to include renewable fuels, food and other substances. But he doesn't see this happening soon.

“This is a proof-of-concept paper,” agrees Erb. “It will take a couple years until we see this organism applied.”

STEVE GSCHMEISSNER/SPL

## The Best Antibody Discovery Technology Is Now at Your Fingertips

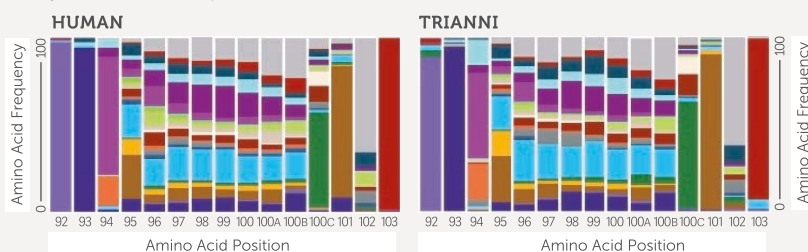
### Trianni Mouse Antibodies are a Match for Humans

The Trianni Mouse™ platform is a transgenic antibody discovery platform offering the entire human variable gene diversity in a single organism.

The V(D)J gene segments in The Trianni Mouse are chimeric, but the variable domains of **antibodies made by the mouse are entirely human**. The result is human antibody leads generated from antibody genes optimized for function in the mouse.

To learn more about this innovative platform, visit [Trianni.com](http://Trianni.com).

Heavy Chain CDR3 Compositions of a Human Individual and a Trianni Mouse are Almost Identical



Different amino acids are represented by different colors

# TRIANNI

Exceptional Human Antibody Discovery Technology



# GENE THERAPY'S TOUGHEST TEST

As a beleaguered field gains momentum against genetic disorders, sickle-cell disease looms as one of its biggest challenges. **By Heidi Ledford**



For years, Grajevis Bakatunkanda's sickle-cell anaemia went undiagnosed.

**G**rajevis Bakatunkanda's mother knew the signs: when her son lost interest in dinner, that meant the pain was on its way. It would strike, like clockwork, nearly every week. Soon the shy, skinny boy would be at the hospital near their home in the Democratic Republic of the Congo, where doctors would provide morphine for the pain and invariably diagnose him with malaria.

It turns out the doctors were wrong. The culprits were not parasites, but Bakatunkanda's own red blood cells. Normally soft and springy, some of the boy's cells were becoming deformed and stiff, like splinters of wood. They would lodge in his capillaries, choking the blood flow to vital organs and sending waves of crushing pain into his back and chest.

It wasn't until the family immigrated to Cape Town, South Africa, in 2003, that they learned Bakatunkanda had sickle-cell anaemia, one of the world's most prevalent genetic disorders, and one that has been studied for more than a century. But the diagnosis did little to ease the boy's pain: the cocktail of drugs that he was prescribed – each of them in use for more than half a century and none developed specifically for sickle-cell disease – failed to break the cycle.

Now, Bakatunkanda is 22, and modern solutions are on the horizon in the form of gene therapies. After decades of work and some painful setbacks, techniques that involve altering a person's genome have begun to win approval for a handful of rare disorders. Scientists are now working to extend the latest advances – including some that use newer gene-editing technologies – to sickle-cell disease, a condition that affects some 20 million people worldwide (see 'How to stop sickling'). There are more than half a dozen active clinical trials, and more are planned. "The studies are just literally coming back to back now," says Lakshmanan Krishnamurti, a paediatric haematologist at Emory University in Atlanta, Georgia. "It's a very exciting time."

But sickle-cell disease could challenge the gene-therapy field both ethically and technologically. Gene therapies that have been approved for other conditions have come with price tags in excess of US\$1 million. But sickle-cell disease is concentrated in regions of the world such as sub-Saharan Africa, India and the Caribbean, where few have the resources to foot such a hefty bill. The experimental treatments for sickle cell are also complex, requiring long hospital stays and the expertise of large academic medical centres. Even for people who can access such resources, the risks might not always be worth it.

As data drift in from early trials, scientists are working to improve their approaches, and funders have already begun to tackle the equity question. On 23 October, the US

AURÉLIE MARRIER/D'UNIVILLE FOR NATURE



National Institutes of Health (NIH) and the Bill & Melinda Gates Foundation announced that they would invest at least \$200 million over the next four years to bring gene-based treatments for sickle-cell disease and HIV to low-resource settings.

Bakatunkanda, who founded a support group for people with sickle-cell disease, is confident that gene therapy, if shown to be effective, will one day reach his country, despite its high cost and daunting complexity. “Definitely it will,” he says. “Because South Africa is rising.”

He and others must keep a tight rein on their expectations. “My patients with Internet access, now they are coming to me: ‘Can we go for gene therapy?’” says Dipty Jain, a paediatrician at the Government Medical College in Nagpur, India. “I advise them, ‘This is not yet for you.’”

## A medical revolution

The elongated, oddly shaped blood cells typical of sickle-cell disease were first noted in 1910 in a young dental student from Grenada, West Indies, named Walter Clement Noel<sup>1</sup>. Forty years later, the underpinnings of the disease began to come into view, when biochemist Linus Pauling and his colleagues reported that changes in the structure of haemoglobin, the oxygen-carrying protein found in red blood cells, were altering the shape of the cells<sup>2</sup>.

The publication marked the first time the effects of a genetic disorder had been traced to their molecular roots. Pauling dubbed the condition “a molecular disease”. Some years later, researchers identified changes in the  $\beta$ -globin protein as responsible<sup>3</sup>. A mutation in both copies of the gene encoding for this protein results in disease; a single mutated copy correlates with few symptoms and protects the bearer from blood-dwelling parasites such as those causing malaria. This, in part, is why the disease exists in relatively high rates where malaria is endemic. “This is really the basis from where everything that we know today on human medical genetics has been developed,” says Ambrose Wonkam, a geneticist at the University of Cape Town.

Seventy years after Pauling’s discovery, sickle-cell disease is still underdiagnosed in many African countries, says haematologist Olu Akinyanju, the founder and first chairperson of the Sickle Cell Foundation Nigeria in Lagos. Yet early diagnosis can save lives. More than 300,000 people are born with the disease each year, and without prophylactic antibiotics and vaccines to help ward off other infections, most will die before the age of five. Those who survive face a lifetime of risk for pain crises, stroke and infection.

Sickle cell disease’s close association with low-income countries has meant that it has historically received little attention from

pharmaceutical companies and governments in richer regions. Many African nations have such pressing public-health needs that it has been difficult to push sickle-cell disease to the top of their priority lists, says Akinyanju, who has campaigned for decades to get African governments to establish treatment plans.

**“IF WE REFINE THE TECHNOLOGY, IT WILL BE AFFORDABLE IN THE LONG RUN.”**

Over the past ten years, however, Akinyanju and others have noticed a shift. As advocates and clinicians push for newborn screening and early intervention, people with sickle-cell disease have begun living longer. The condition is not as stigmatized as it once was. Akinyanju proudly ticks off friends with sickle-cell disease who have lived into their 60s and beyond, becoming doctors, judges and world travellers.

The World Health Organization and the American Society of Haematology have also

worked to bring the disease to the attention of researchers and pharmaceutical companies. And Bakatunkanda and other immigrants have raised awareness in wealthier nations, says Wonkam. There are signs that this attention is paying off. On 25 November, the US Food and Drug Administration approved a drug for sickle-cell that aims to reduce clumping between haemoglobin molecules.

But although gene therapy might seem a rational approach for one of the world’s best-known genetic diseases, the field has faced its setbacks. Early attempts were marred by the high-profile death of Jesse Gelsinger in 1999, who was participating in one of the first gene-therapy clinical trials. A procedure used during the trial to replace immune-system genes in blood stem cells caused leukaemia in several of the participants.

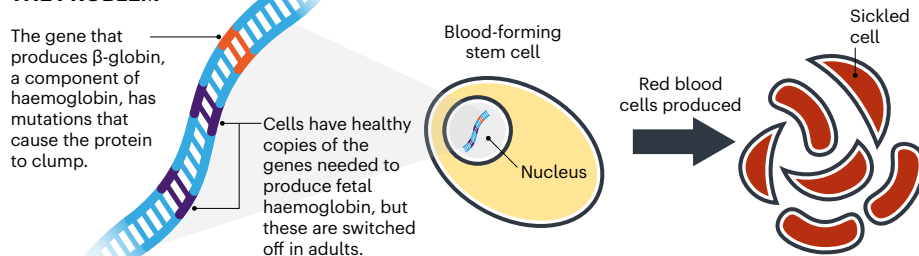
Against that backdrop, some felt that it was premature to apply gene therapy to sickle-cell disease, says haematologist David Williams at Boston Children’s Hospital in Massachusetts. “Sickle cell is not an immediately lethal disease,” he says. “In some ways, it wouldn’t be ethical to treat those patients with a highly risky experimental approach.”

Furthermore, the tools were not yet up to the task, says Donald Kohn, a specialist in paediatric bone-marrow transplants at the University of California, Los Angeles. If researchers were to shuttle in a normal haemoglobin gene, it would need to be able to crank out large amounts of

## HOW TO STOP SICKLING

When blood-forming stem cells have mutations in the gene for  $\beta$ -globin, they produce red blood cells that can become hardened and sickle-shaped. Gene therapies for sickle-cell disease aim to remove these stem cells from the bone marrow, alter their genomes and then replace them in the patient.

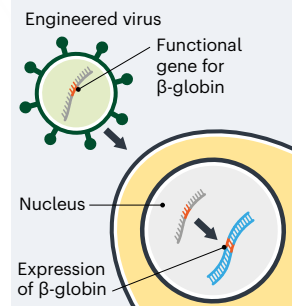
### THE PROBLEM



### THE SOLUTIONS

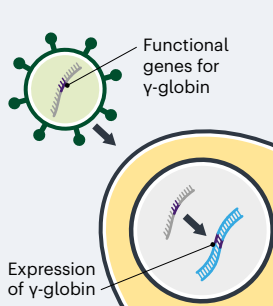
#### $\beta$ -GLOBIN RESCUE

Scientists engineer a virus to deliver a functional copy of the gene that produces  $\beta$ -globin. The gene has been modified to prevent sickling.



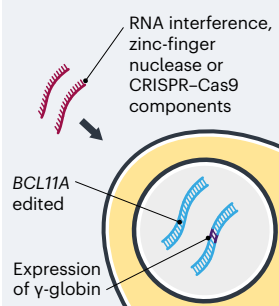
#### $\gamma$ -GLOBIN ADDITION

A virus delivers genes that produce  $\gamma$ -globin, a component of fetal haemoglobin. The genes are modified to remain switched on in adult cells.



#### BREAKING THE OFF SWITCH

Various gene-silencing and editing technologies turn off production of BCL11A, a protein that normally prevents expression of  $\gamma$ -globin genes.







AURÉLIE MARRIER D'UNIVILLE FOR NATURE

**Bakatunkanda takes a cocktail of older drugs to fight his disease. Treatment with gene-therapy sounds attractive, but he knows there are risks.**

protein to sufficiently mute the effects of the sickled version. Early gene-therapy technologies were not able to express genes in human cells at such high levels, says Kohn.

But despite the setbacks, some gene-therapy researchers pushed on, developing safer and more potent ways to shuttle genes into cells. They broke through in 2016, when the European Commission approved a gene therapy for treating ADA-SCID, a rare immune disorder that often kills children before their first birthday. Then in 2017, the US Food and Drug Administration approved a gene therapy to treat a rare form of blindness.

By this time, some researchers had turned their attention back to sickle cell, armed with improved tools and with the backing of the biotechnology industry. Current trials are taking a variety of approaches. Kohn is trying to insert a copy of the  $\beta$ -globin gene that has been modified to resist sickling. So is Bluebird Bio, a company in Cambridge, Massachusetts. The firm looks set to be the first to win approval to market such a treatment in the United States, according to Yaron Werber, a biotechnology analyst at Cowen, a financial-services company in New York City.

Others are introducing modified copies of the genes that encode fetal haemoglobin, a form of the protein that is produced in the

developing fetus but usually shuts off soon after birth. Fetal haemoglobin is an attractive option because it works about as well as the adult version, and it prevents defective haemoglobins from clumping together.

A third approach seeks to block a mechanism that switches off production of fetal haemoglobin after birth. The usual off-switch is a protein called BCL11A, and suppressing it in mice with sickle-cell disease can keep fetal-haemoglobin levels high well into adulthood and prevent symptoms of the disease<sup>4</sup>. In Boston, Williams has licensed technology to Bluebird Bio that uses a technique called RNA interference to dial down expression of the gene encoding BCL11A in blood stem cells. Sangamo Therapeutics in Richmond, California, in partnership with Sanofi in Paris, is using gene-editing tools called zinc-finger nucleases to create mutations that disable the gene. And Vertex Pharmaceuticals in Boston has teamed up with CRISPR Therapeutics in Cambridge, Massachusetts, to do much the same using the CRISPR–Cas9 gene-editing technique. In all three approaches, blood-producing stem cells are removed from the body, genetically altered – often with the help of a virus – and then reintroduced into the bone marrow. Before the cells are replaced, participants are typically treated with a chemotherapy called busulfan

to destroy the remaining diseased stem cells and help the reintroduced, genetically altered cells to take over.

That kind of regimen is risky: participants can develop acute and severe anaemia. The treatment wipes out their white blood cells, and wreaks havoc on the lining of the gut, potentially leaving them dependent on intravenous nutrition. Many will need to stay in the hospital for more than a month. The chemotherapy also causes infertility, and can cause cancer later in life.

This means that gene therapy would probably be used only in those with the most serious forms of sickle-cell disease. Yet many of those people will also have heart, kidney or liver damage that would make the chemotherapy too dangerous.

Sickle-cell disease complicates the therapy in other ways, too. In many cases, when doctors harvest bone marrow, patients first receive a drug that makes it easier to collect blood stem cells. But that is too dangerous to use in people with sickle-cell disease because it raises the risk of pain crises. And because diseased red blood cells die faster than healthy ones, the stem cells in a person with sickle-cell disease must work harder to produce new blood cells. This can leave them in poor condition for harvest and growth in laboratory cultures. As a result,

participants often need blood transfusions just before harvest to ease the stress on their stem cells. Despite these challenges, early signs of success have been making headlines. One of the men in Williams's RNA-interference trial has been symptom-free for one year. And the first patient in the CRISPR trial has now left the hospital after completing the gruelling therapy. On 19 November, Vertex and CRISPR Therapeutics announced that the person has not experienced any pain crises and has maintained a high level of fetal haemoglobin for four months. Both trials have generated excitement on social media – too much, in some cases. “I have difficulty right now with folks being excited about the discharge of a patient from the hospital, as if that were tantamount to a cure,” says Alexis Thompson, a haematologist at Northwestern University in Chicago, Illinois. “That’s a pretty low bar.”

Still, there is cause for cautious optimism. So far, none of these trials has been stopped for safety concerns. And Bluebird Bio has treated 13 people, some of whom have been monitored for a year after treatment with no severe pain crises, the company reported in June. The gene therapy used was approved in the European Union in June to treat some people with a related genetic blood disorder called  $\beta$ -thalassaemia.

But a major concern for many people is cost. The treatment for  $\beta$ -thalassaemia runs to roughly \$1.8 million – not including the hospital stay and other associated costs.

This is still potentially cheaper than standard treatments over the course of a lifetime, says Mani Foroohar, an analyst at the investment bank SVB Leerink in Boston, Massachusetts. Also, Bluebird Bio has established an unusual fee structure: payments are made over the course of five years, and can be halted if the treatment stops working. Still, Foroohar says, it’s not clear whether the same model will be possible in other regions.

The price tag is certainly well beyond the means of many of Jain’s patients in central India, who come to her hospital because they can’t otherwise afford the roughly \$3 per month that it costs for standard treatments. Even in the United States, access to the gene therapies is likely to be a challenge. This is particularly true for Black Americans, who tend to have more limited access to health care than White Americans. Although the trials are still in their early days, Krishnamurti urges interested people to begin advocating immediately for access to the therapies. “It’s an enormous ethical issue,” says Krishnamurti, who counsels people with sickle-cell disease each week from his hospital in Atlanta. “In my community conversations, I say, ‘You had better be at the table, otherwise these decisions will be made without your input.’”

At the Cincinnati Children’s Hospital in Ohio, haematologist Punam Malik is hoping

to take the first steps towards making gene therapy cheaper and simpler. Malik trained as a doctor in India, where she saw many people with sickle-cell disease and related conditions. When she immigrated to the United States about 30 years ago, she vowed to make sure that her research would benefit people in resource-poor countries.



**IT’S A GREAT LOFTY GOAL. I THINK THE SCIENCE IS ADVANCING PRETTY RAPIDLY.”**

Now, Malik is leading a trial that introduces stem cells that produce fetal haemoglobin. It uses low doses of a drug called melphalan to remove diseased cells from the bone marrow, which should make the treatment less toxic than the usual busulfan. Her hope is that the technique will reduce the need for a long hospital stay, making the treatment cheaper, safer and more practical.

But the approach has been criticized by others, who worry that the low-dose approach might leave behind some uncorrected cells, and make the therapy less effective. “You want to do as well as you can,” says Stuart Orkin, who studies blood disorders at Boston Children’s Hospital.

Malik counters that once a high dose has been established as effective, it is hard to scale it back. She points to the example of cancer chemotherapy: in some cancers, researchers are reducing the dose of some drugs and finding that they work just as well as, if not better than, the higher doses tried initially. But it has taken oncologists decades to take that step, she notes. “I might fall flat on my face, and I might have to dial up. But it will be very difficult for the others to dial down,” she says.

Her trial has also run up against the practical realities of exporting gene therapies to regions with fewer resources. Her team received FDA approval to carry out the trial only at Cincinnati Children’s Hospital. But after Malik gave a talk at a conference in Jamaica, someone with sickle-cell disease approached her asking for help and describing multiple hospital visits for pain crises.

So, Malik developed a collaboration in Jamaica. “I felt we had to,” she says. It took the team about two years to get the necessary approvals and funding. And then the clinical team in Jamaica ran up against another problem: lack of reliable blood for transfusion.

The team reported in April that its first

patient has experienced only two pain crises in the 18 months since treatment ended and has maintained high levels of haemoglobin. The team has since treated a second person, and two more are lined up to take part, Malik says.

The trouble is not only the expense and practicalities, but also the availability of clinicians and facilities who can handle stem-cell transplants. Rural regions already struggle to supply people with with hydroxyurea, a relatively cheap medicine that reduces the rate of pain crises. It’s hard to imagine these regions having enough personnel to monitor recipients of gene therapy over the long term, says anthropologist Duana Fullwiley at Stanford University in California.

Some argue that it is too early to think about such issues. “If we refine the technology, it will be affordable in the long run,” says Wonkam. “The price right now for me is not the problem. The focus needs to be on the efficiency.”

But others think that the time to start thinking about global access is now. To do otherwise “would be almost unethical”, says NIH director Francis Collins.

Collins thinks that the key to fulfilling the NIH’s project with the Gates foundation will be in finding ways to deliver the corrected genes or gene-editing tools to bone-marrow stem cells that don’t involve having to remove the cells first, making therapies cheaper and easier to deliver. It is an ambitious goal – and one that is occasionally met with scepticism, Collins says. “Sometimes there was a vague sense of, ‘Boy, you’re just outside the boundaries of reality there, Collins,’” he says.

There are already suggestions that the viruses typically used to shuttle genes into cells in a dish can be modified to insert genes into blood-producing stem cells while they’re still in the body, notes Kohn. “It’s a great lofty goal,” he says. “I think the science is advancing pretty rapidly.”

For Bakatunkanda, his salvation turned out to be ageing, not medicine. Some people with sickle-cell disease fare worse as children than as adults, he says, and he thinks he is one of them. He still has crises, but not nearly as often. In recent months, he has taken on activities such as hiking and bodybuilding that he once thought were off-limits. “I just know how far I can push myself,” he says.

But he would prefer a life without the constant threat of pain crises and strokes. He is aware of the promise of gene therapies, but knows that it is not yet clear whether they will provide a cure. “I would prefer that,” he says. “But at the moment it’s not a guarantee.”

**Heidi Ledford** writes for *Nature* from London.

1. Herrick, J. B. *Arch. Intern. Med.* **6**, 517–521 (1910).
2. Pauling, L., Itano, H. A., Singer, S. J. & Wells, I. C. *Science* **110**, 543–548 (1949).
3. Ingram, V. M. *Nature* **178**, 792–794 (1956).
4. Xu, J. et al. *Science* **334**, 993–996 (2011).





ARTERBA/UNIVERSAL IMAGES GROUP VIA GETTY

Triassic rocks in the Italian Dolomites bear evidence of a surprisingly wet episode in Earth's history.

# A MILLION YEARS OF TRIASSIC RAIN

An extended bout of warm, wet weather 232 million years ago might have triggered the rise of the dinosaurs and completely altered the history of life on Earth. **By Michael Marshall**

**A**lastair Ruffell could see there was something odd about the rocks near his childhood home in Somerset, UK. The deposits hail from the Triassic period, more than 200 million years ago, and most are a dull orange-red, signifying that they formed when the region was a parched landscape, baked by the sun. Nothing strange there. But outcrops on Somerset's Lipe Hill have a thin stripe of grey

running through the heart of the red stone. That band signals a time when arid desert disappeared and the region transformed into a swampy wetland. For some reason, an incredibly dry climate had turned wet, and stayed that way for more than a million years.

The change intrigued Ruffell when he first found the outcrops in the mid-1980s, but the young geologist had a PhD project to finish. So he put the Triassic puzzle to one side, until a chance encounter in 1987 with another young

scientist, palaeontologist Michael Simms. During his postdoctoral studies, Simms had discovered evidence of extinctions in the Late Triassic, during Ruffell's mysterious wet period. In the late 1980s, the pair pushed the idea that the two findings were connected, but for years, their results were dismissed.

Three decades later, there is a growing consensus that they were right, after all. Something strange happened in the Late Triassic — and not just in Somerset. About

232 million years ago, during a span known as the Carnian age, it rained almost everywhere. After millions of years of dry climates, Earth entered a wet period lasting one million to two million years. Nearly any place where geologists find rocks of that age, there are signs of wet weather. This so-called Carnian pluvial episode coincides with some massive evolutionary shifts.

Perhaps most dramatically, the Carnian pluvial might have overlapped with when a rare group of reptiles – early dinosaurs – evolved into a diverse group and came to dominate land ecosystems. The Carnian could have paved the way for the spectacular dinosaurs that evolved later, including *Stegosaurus* and *Tyrannosaurus*.

Other groups also left the Carnian in very different shape from how they had entered it: reef-building corals and marine plankton were all becoming more ‘modern’ – moving evolutionarily closer to the forms alive today. The period could even have seen the appearance of the first mammals. “It was almost like a turning point between some elements of a more ancient world and a modern world,” says Simms.

After years in obscurity, the Carnian pluvial is becoming a major research focus. In May 2017, scientists gathered for the first conference dedicated to the period, held at the Institute for Advanced Study in Delmenhorst, Germany. Since then, the *Journal of the Geological Society* has dedicated two special issues to the topic. Over the past decade, many researchers have begun studying Carnian rocks intensively. They want to understand why the climate changed, and why that led to such dramatic evolutionary shifts. Evidence now points to massive volcanic eruptions.

This is a remarkable turnaround for an event discovered almost by accident in the 1980s.

## A chance encounter

It began when Simms, now at National Museums Northern Ireland in Holywood, went to the University of Liverpool, UK, for a postdoctoral research fellowship. He studied crinoids: marine animals, related to starfish, that resemble flowers or feathers.

Simms focused on crinoids living in the Triassic period, which ran from 252 million years ago to 201 million years ago. The Triassic was bookended by two of the most troubled times in Earth’s history: it started right after a mass extinction at the close of the Permian period, and ended with another mass extinction at the Triassic–Jurassic junction.

But Simms was in for a surprise. “By the tail end of 1987, it had become clear that there was a quite significant extinction among the Triassic crinoids,” he says. But the die-off came tens of millions of years before the end of the period. This placed the extinction in the Carnian: the fifth of seven shorter ages in the Triassic.

Intrigued, Simms returned to the University of Birmingham, UK, where he had done his doctorate, for a visit. His old office was occupied by palaeontologist Paul Wignall and Ruffell, who is now at Queen’s University Belfast, UK.

Ruffell’s studies focused on sediments from the later Cretaceous period, but for fun he was investigating Triassic rocks called the Mercia Mudstone Group, which mostly reflect dry climates. It was in the Carnian section of these rocks that he found a thin layer of grey sandstone, rich in fossils such as sharks’ teeth. It was the remains of a river or delta. “Slap-bang



**I REMEMBER ONE OR TWO QUITE SENIOR ACADEMICS THOUGHT IT WAS A PREPOSTEROUS IDEA.”**

in the middle of all this horrible arid stuff was this probably rather pleasant environment,” says Ruffell.

When the three were chatting, Simms mentioned the Carnian crinoid extinctions. According to Simms and Wignall, Ruffell replied: “It was raining then. Perhaps the crinoids didn’t like the rain.” It was a flippant remark, one Ruffell does not remember making, but it struck a chord with Simms. Changing climates can cause extinctions, so perhaps this was the case for the shifts in the Carnian.

Simms and Ruffell began investigating, and found that there was also evidence of a wet spell in the Carnian in rocks from Germany, the United States, the Himalayas and other places. What’s more, it was not just the crinoids that faced extinctions: amphibians and land plants lost members, too. In 1989, the pair published evidence for an event they named the Carnian pluvial episode<sup>1</sup>, using the geological term referring to rain.

The results didn’t make much of a splash, except from some researchers who attacked the idea. “I remember one or two quite senior academics thought it was a preposterous idea,” says Simms.

In 1994, a team led by Henk Visscher of Utrecht University in the Netherlands published a strongly worded rebuttal claiming that although some spots might have grown rainy during this time, many environments remained dry<sup>2</sup>. Visscher argued that, instead of an increase in rainfall, the evidence could be explained away by “high groundwater tables”.

Rebuffed, Simms and Ruffell changed course. “We just moved on to all sorts of other

things,” says Simms. Whereas Simms pursued a career in geology and palaeontology, Ruffell became an expert in forensic geology.

## Wet world

However, the Carnian pluvial episode did not go away. Geologists in Europe, especially Italy, continued amassing evidence for wet conditions around 232 million years ago. The coincidences piled up.

In the Late Triassic, the world looked nothing like today’s (see ‘Time for a change’). The landmasses were all connected, forming a ‘super-continent’ called Pangaea, where the climate was hot and dry, especially in the interior. Land ecosystems were dominated by reptiles, including the first dinosaurs. There were no flowers, grasses or birds.

There were also no mammals, but the Carnian might have been when that changed. In 2005, P. M. Datta at the Geological Survey of India in Kolkata described a single mammal tooth from Carnian-aged rocks in India<sup>3</sup>. Another tooth, discovered in Carnian rocks in Germany, might also have belonged to a mammal<sup>4</sup>.

The origin of mammals is a topic that triggers strong debate. Wignall, who is now at the University of Leeds, UK, says they could have appeared during the Carnian, but it’s also possible that there are earlier ones we have not found yet. And many palaeontologists argue that true mammals did not emerge until the Jurassic, millions of years later. If so, the Carnian fossils are not from mammals, although they could be from their ancestors.

Whatever the case is with mammals, a string of discoveries in the past decade or so offers strong evidence for other evolutionary shifts. Researchers reported in 2013 that the Carnian saw the origin of marine organisms called calcareous nannoplankton<sup>5</sup>. These single-celled organisms surround themselves with hard shells of calcium carbonate. Today, they form huge blooms and are known as the ‘grass of the sea’. They have a major role in cycling carbon between the air and the ocean.

Another group that underwent major changes in the Carnian was the scleractinian corals, which build today’s giant coral reefs. Scleractinians emerged earlier in the Triassic, but it was not until the Carnian or shortly afterwards that they began constructing big reefs. Isotopic evidence and other clues in fossil corals suggest this could be when they acquired their modern symbionts: photosynthetic algae that supply them with nutrients<sup>6</sup>.

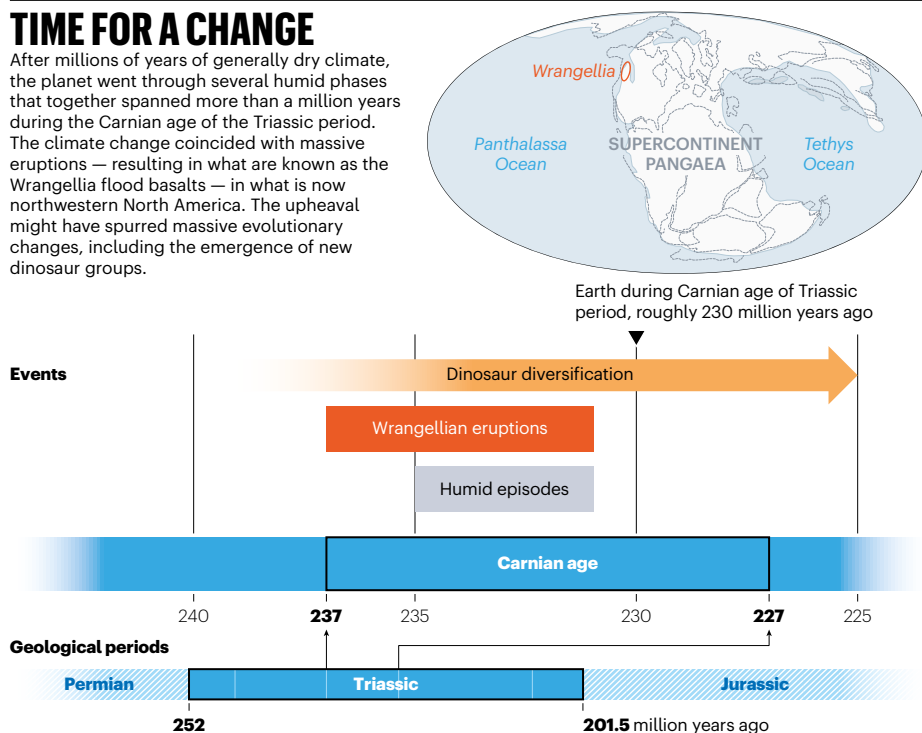
## A fiery time

Not everyone is convinced that the world went through a warm, wet phase in the Carnian. “Still, I have my doubts,” says Visscher. He accepts that the climate changed, but says rainfall could have become more seasonal, leading to annual blooms of vegetation. Similarly, Matthias Franz at the Georg-August



## TIME FOR A CHANGE

After millions of years of generally dry climate, the planet went through several humid phases that together spanned more than a million years during the Carnian age of the Triassic period. The climate change coincided with massive eruptions — resulting in what are known as the Wrangellia flood basalts — in what is now northwestern North America. The upheaval might have spurred massive evolutionary changes, including the emergence of new dinosaur groups.



University of Göttingen, Germany, has found evidence that the extra damp might have been caused by rising seas<sup>7</sup>, at least in parts of Europe, although it is not clear that this can account for the changes elsewhere. Still, Franz emphasizes that the period is significant anyway. “There is obviously something happening at this time,” he says. “The question is what.”

Simms and Ruffell had previously suggested that volcanic eruptions were responsible for the climate change, and geologists knew there was a prime candidate: the cataclysm that created massive basalt formations — several kilometres thick in places — running from British Columbia in Canada to Alaska.

Dubbed the Wrangellia Terrane, after Alaska’s Wrangell Mountains, these lavas are part of a Large Igneous Province formed by volcanoes spewing out huge volumes of lava over hundreds of thousands or millions of years, around 232 million years ago. The volcanoes were submarine, but emerged above the water as lava continued to pour out, says Andrea Marzoli at the University of Padua in Italy.

If these vast eruptions happened at the same time as the Carnian pluvial episode, they could have released enough carbon dioxide to warm the globe. And that could have increased rain, by enhancing evaporation from seas and rivers. Some scientists have come to regard the name Carnian pluvial as misleading, because the main change at the time would have been an episode of global warming.

“The natural thing to do was to understand if this increase in rainfall, that was seen everywhere, was triggered by injection of CO<sub>2</sub> in the atmosphere,” says Jacopo Dal Corso, a geologist at the University of Leeds. His team analysed samples of carbon-rich Carnian material

from the Italian Alps. In 2012, the researchers reported unusually low levels of carbon-13, a heavy isotope of carbon, during the Carnian pluvial<sup>8</sup>. This indicated that a huge volume of the lighter isotope, carbon-12, was injected into the atmosphere — and eruptions in Wrangellia could have been the prime source.

Subsequent studies have backed Dal Corso’s

**“ONE OF THE FASCINATING THINGS ABOUT THIS INTERVAL IS HOW MANY MODERN GROUPS APPEAR.”**

claim that the carbon cycle was perturbed during the Carnian for about one million years<sup>9</sup>, owing to the eruptions. But for some, the link remains tentative because uncertainty in the dating of rocks makes it hard to definitively say that the Wrangellia eruptions happened at the same time as the climatic and evolutionary changes in the Carnian. Wignall says this is because the Carnian has not yet been studied intensively; uncertainties can span one million years. Marzoli plans to sample Wrangellia next summer, partly to clarify its age. According to him, Wrangellia is the most likely explanation because there are no other candidates.

Meanwhile, the list of evolutionary changes that happened in the Carnian pluvial continues to grow.

The most dramatic claim is that the Carnian was crucial for the dinosaurs’ rapid evolutionary expansion. Evidence indicates that dinosaurs emerged before the Carnian, about 245 million years ago, but those earliest creatures are very rare and only a few species are known.

What’s clear is that dinosaurs changed drastically. At the start of the Carnian, they were all small and bipedal. But by the end, the two major groups had emerged. These were the ornithischians, which later included *Stegosaurus* and *Triceratops*; and the saurischians, which gave rise to huge, long-necked species such as *Brachiosaurus*, and theropods such as *Tyrannosaurus rex* and birds. Mike Benton, a palaeontologist at the University of Bristol, UK, and his colleagues documented some of these changes by using well-dated samples from the Alps to create a high-resolution timescale of animal tracks in the Late Triassic<sup>10</sup>. The early Carnian was dominated by reptiles called crurotarsans. But by the end of the Carnian, the dinosaurs dominated. This shift took just 4 million years, and coincided with the pluvial episode. And after that rapid rise, dinosaurs ruled the world for more than 150 million years.

With all these changes happening, and the fuzzy dating of rocks from the Carnian, researchers are struggling to create a coherent picture of how the climate changed and how that affected ecosystems. But the Carnian has become a hot topic. “One of the fascinating things about this interval is how many modern groups appear, from vertebrates all the way down to plankton,” says Wignall.

This was one of life’s major transitions. The planet was still recovering from the end-Permian extinctions and the Carnian saw the rise of groups that have ruled the world ever since.

The two researchers who started this whole affair are surprised and delighted by what has happened. Simms is content to watch from the sidelines, but Ruffell has resumed studying Carnian geology. The irony, Ruffell says, is that his Carnian studies were only a hobby. This dramatic period that shook up evolutionary history was found, he says, by “a couple of guys who really shouldn’t have been working on it in the first place”.

**Michael Marshall** is a science journalist in Devon, UK.

1. Simms, M. J. & Ruffell, A. H. *Geology* **17**, 265–268 (1989).
2. Visscher, H. et al. *Rev. Palaeobot. Palynol.* **83**, 217–226 (1994).
3. Datta, P. M. J. *Vert. Paleontol.* **25**, 200–207 (2005).
4. Lucas, S. G., Heckert, A. B., Harris, J. D., Seegis, D. & Wild, R. J. *Vert. Paleontol.* **21**, 397–399 (2010).
5. Preto, N., Willems, H., Guaiumi, C. & Westphal, H. *Facies* **59**, 891–914 (2013).
6. Stanley, G. D. Jr *Science* **312**, 857–858 (2006).
7. Franz, M. et al. *Glob. Planet. Change* **122**, 305–329 (2014).
8. Dal Corso, J. et al. *Geology* **40**, 79–82 (2012).
9. Dal Corso, J. et al. *Earth Sci. Rev.* **185**, 732–750 (2018).
10. Bernardi, M., Gianolla, P., Petti, F. M., Mietto, P. & Benton, M. J. *Nature Commun.* **9**, 1499 (2018).

SOURCE: A. RUFFELL, J. DAL CORSO & M. BENTON *GEOSCIENTIST* **28**, 10–15 (2018).

# Books & arts

## Why pipelines persist amid geopolitical turmoil

In a new book, Thane Gustafson analyses the Russia–Europe gas trade. **By Andrew Moravcsik**

**M**any people imagine that geopolitics drives the energy trade between Russia and Europe. As the story goes, each side seeks to exploit gas and oil to influence the other in the big game of power politics – and Russia seems to have the upper hand. The European Union now imports nearly 40% of its natural gas from Russia. For decades, national-security specialists have recommended that Europeans reduce their dependence on these imports at any cost. Most recently, a fierce debate over Nord Stream 2 – a second Russian pipeline across the Baltic Sea to Germany – has led US congresspeople to threaten sanctions.

Political scientist Thane Gustafson challenges this view in *The Bridge*. He argues that the trade in gas reflects slow-moving patterns of market demand and supply, which in turn stem from incremental changes in technology for pumping, piping and consuming fuel. The result is a pattern of remarkably stable economic interdependence that seems impervious to the geopolitical environment.

As extraction and pipeline technology opened up Soviet gas fields in the 1960s, and the ongoing postwar reconstruction of Europe stoked demand, East–West gas trade became all but inevitable. Ever since, Russia has wanted to supply gas and Europe has wanted to buy it. The past 50 years have seen energy shocks and gluts; major political crises from Poland to the former Yugoslavia; the fall of the Soviet Union and rise of Russian President Vladimir Putin's authoritarian state; outright warfare in Ukraine and elsewhere; massive experiments in deregulation; and the rise of environmentalism. Yet relations between Europe and Russia in the natural-gas sector have remained nearly constant. This is because change is slow in three factors: proven reserves of gas, aggregate demand for energy and investment in physical infrastructure to link the two.

*The Bridge* is an overview rather than a work of original research. Yet it offers a readable, intelligent, even-handed historical

interpretation of this modern economic relationship. It divides East–West natural-gas relations neatly into three distinct periods.

The first begins around 1960, with the spread of transport and use of natural gas in Europe, originally limited to small local networks in Italy and the Netherlands. Backed by US expertise, Europeans began to consider long-distance gas pipelines from Siberia, and made Western industrial equipment, investment and technical know-how available to the Soviet Union. The rigidity of the Communist system meant that production took almost a decade to come online. Eventually, the gas arrived, at first flowing through a terminal in Austria.

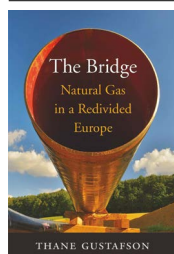
The second period begins after 1970, when the quantity of Russian natural gas entering Europe increased. European consumption expanded quickly; gas proved cheaper and environmentally cleaner than coal or oil. Other countries, notably undersea-gas producers Norway and Britain, also created highly centralized systems for exploiting and piping the fuel. Yet the vast, low-cost Russian reserves enjoyed a comparative advantage, rising to provide almost half of consumption in European countries, prominent among them Germany and Italy.

This period, Gustafson argues, demonstrates the exceptionally stable nature of this type of international economic cooperation. Pipelines take decades to build, then tend to operate for decades more, often governed by just one or two long-term contracts. The physical, tangible linkage between



producer and consumer “automatically creates a mutual dependence”, he writes. Moreover, because pipelines are centralized, they encourage domination of the market by monopolies – in the 1970s, these consisted of the Soviet Ministry of the Gas Industry, and European national or regional utilities. Natural gas, or anything else that travels through a fixed infrastructure, becomes a “relationship commodity”: investments, personal contacts and market shares follow the technology.

This, Gustafson avers, is why the East–West gas trade has remained impervious to geopolitical disruption. In 1968, shortly after the Soviet invasion of Czechoslovakia, Austria accepted the first Russian gas shipments into Europe. In 1981, when the pro-democracy Solidarity movement in Poland led to the Soviet-backed imposition of martial law, the US administration under president Ronald Reagan imposed sanctions on exports of pipeline technology. It could afford to do this because it was largely uninvolved in the



**The Bridge: Natural Gas in a Redivided Europe**

Thane Gustafson  
Harvard Univ. Press  
(2020)

ALEXANDER NEMENOV/AFP/GETTY





Russian natural-gas pipelines in northwestern Siberia.

East–West energy trade.

Yet behind the scenes of these political upheavals, the real stakeholders acted differently. The Soviet Union developed home-grown alternative compressor and pipe technology – crucial for transporting gas – and Europe continued to sell technology that the Soviets could not produce at home.

The third period began around 1990. Geopolitics grew more unruly. The Soviet Union collapsed in 1991. The gas ministry was turned into the massive state-owned corporation Gazprom, which was then largely privatized. Putin, who became president in 2000, brought Gazprom back under near-total state control. Russia also provoked a series of interventions and conflicts in Georgia, Moldova, Syria and Ukraine. The West responded by imposing sanctions – limits on investment and exports in sensitive military and civilian technologies, and even on energy investment. Russia's countersanctions largely targeted Western agricultural exports. More recently,

Russia has become involved in the disruption of elections in the West, and in cyberwarfare. Yet gas quietly continues to flow through the East–West pipeline.

Much of the book's analysis of the most recent period focuses on another potentially disruptive change: new EU regulations. Gustafson makes much of the fact that, 30 years ago, the European Commission began pressing to open up the European energy market to greater competition. Directives render prices more transparent and uniform, and compel firms to supply gas across borders. At the same time, the commission is acting more forcefully to limit monopolies and cartels, and domestic deregulation has led to the rise of new corporate players.

Overall, this concerted EU policy has further strengthened Europe's hand. Russia cannot use embargoes or market segmentation to exploit individual countries. And Gazprom – which still has a near-monopoly on Russian exports, even though it is losing domestic market share

– cannot acquire dominant positions in Europe. This is a significant development – and, from a Western perspective, a positive one.

Yet it is difficult to discern how EU policies have altered Russia's gas trade with Europe in any fundamental way. Exporting and importing nations alike have found ways to maintain overall control of their markets. If anything, Gustafson's analysis would seem to show that the primary impact of EU consolidation has been to insulate a mutually beneficial economic status quo from disruption.

Gustafson ends by considering long-term threats, which he introduces only to dismiss. For 20 years, conflict with Ukraine – first

**“Russia has become involved in the disruption of elections in the West, yet gas quietly continues to flow.”**

over energy pricing, then over politics – has led Russia to propose new pipelines that geographically circumvent its neighbour. Many worry that new lines, such as Nord Stream 2, might cut Ukraine out entirely. Yet Gustafson remains confident that if this occurs, Kiev, already transitioning away from Russian natural gas, will find new suppliers.

Another threat comes from new technological options for transporting fuel as liquid natural gas, a more fungible form that would permit US imports to Europe. This might create an alternative to stable pipeline politics, although the transition would be slow because of the higher cost of the technology. Also, environmental-protection and climate-change concerns will continue to rise, reducing European demand in the long term. Yet, in the interim, natural gas will remain abundantly available, relatively inexpensive and still environmentally superior to oil, coal or nuclear power.

Gustafson's overall conclusion is thus that Russian gas is likely to remain Europe's major energy bridge to a future world of renewables. He even sees the next few decades as a “golden age of gas”. This is a soberly optimistic conclusion, not least because it suggests that commercial interests will induce modern countries to transcend ideological and geopolitical differences.

**Andrew Moravcsik** is professor of politics and international affairs, and director of the EU Program, at Princeton University in New Jersey. e-mail: amoravcs@princeton.edu

# Comment

## Women from some minorities get too few talks

Heather L. Ford, Cameron Brick, Margarita Azmitia,  
Karine Blaufuss & Petra Dekens

Researchers from racial and ethnic groups that are under-represented in US geoscience are the least likely to be offered opportunities to speak at the field's biggest meeting.

**B**iases – structural, implicit and explicit – exclude many people from science, technology, engineering and mathematics (STEM) education and employment, and devalue their contributions<sup>1,2</sup>. Most studies focus on bias against women. Few data sets offer enough generalizability or statistical power to evaluate the representation of minority ethnic and racial groups, or to examine intersectionality<sup>3</sup>. The latter describes the interwoven forms of discrimination that affect a person from multiple marginalized groups (such as racism, sexism, classism or ageism), locate them in systems of oppression and limit their upward mobility – as might be experienced by a young African American woman in science in the United States.

We offer just such a data set here.

Presenting at scientific conferences is key to academic career progression. Scientists don't just communicate results; they also develop relationships with collaborators and mentors, and identify job and funding opportunities. Giving a talk confers recognition and prestige, particularly for students and early-career researchers. Despite historical inequities, women are now presenting more at conferences<sup>4,5</sup> and colloquia<sup>6</sup>. These gains are especially visible at conferences that are organized by women or that specifically support early-career participants.

We found that US scientists from minority racial and ethnic populations already

under-represented in science had relatively fewer speaking opportunities at a key scientific conference over a four-year period than their proportion in the sample would predict; the imbalance was most severe for women. This disadvantage for under-represented minority groups held across career stage (see 'Who gets the microphone?').

Our results underscore the pressing need to support minority groups at conferences – as elsewhere in STEM – to advance equity and improve research.

### Data set and methods

The American Geophysical Union (AGU) is an international non-profit scientific association with around 60,000 members in 137 countries. Since 2013, the AGU has collected self-reported demographic data from its membership, including gender, race or ethnicity (for US-based academics only), career stage and birth year.

The AGU Fall Meeting is the world's largest Earth- and space-science conference. The attendance each year from 2014 to 2017 was approximately 24,000–28,000 people. Around 22,000 abstracts are submitted for selection as talks or posters each year; few are rejected (<0.05%). Membership is necessary for submitting, although not for attending the meeting.

Abstracts are submitted to topical sessions. Sessions are proposed and organized, and abstracts vetted, by a group of conveners – academics, industry members, government scientists and others. The primary convener must be an AGU member. There are three tracks by which geoscientists get to present at the meeting – two by submission, one by

**“Giving a talk confers recognition and prestige, particularly for students and early-career researchers.”**



invitation. Authors can submit abstracts to conveners, who decide which will become talks and which posters; or authors can submit abstracts just to give a poster. In addition, session conveners directly invite scientists to speak (strictly, to send in abstracts, which generally results in a talk).

The database of 87,544 accepted abstracts from the meetings between 2014 and 2017 offers a unique opportunity to probe inequities of opportunity between demographic groups<sup>5</sup>. Presentations are approximately 34% talks (about one-third of which are directly invited) and 66% posters.

### Career stage

Of US-based authors, 98% ( $n = 53,247$ ) provided career information. Researchers had verified themselves as students (undergraduates and graduates) or the AGU had calculated career stage from years since highest degree obtained: early career (0–10 years); mid-career





Some scientists opt to present posters, others are assigned them instead of being asked to talk.

(10–25 years); and experienced (late career; more than 25 years). Controlling for career stage is crucial because minority racial and ethnic groups are concentrated in the student and early-career stages (see ‘Fewer seniors’). This is due to both a leaky pipeline in education and professional advancement<sup>7</sup> and the fact that senior groups more strongly bear the imprint of historical biases.

### Race, ethnicity, gender

The AGU recorded self-reported ethnicity and race from US-based authors only ( $n = 54,446$ ). Of these, 71% ( $n = 38,768$ ) reported a category (defined as per the US census, see Supplementary information): White (58%), Asian American (7.3%), Hispanic/Latino (3.9%), African American (1.1%), Native American (0.3%) or Pacific Islander (0.2%). The remainder marked ‘other’ (13%) or ‘prefer not to answer’ (13%), or didn’t respond (2.8%). We did not verify whether Native American respondents were

citizens of tribal nations; we acknowledge that self-reported identity is not the same as tribal citizenship. ‘Other’ could refer to individuals who are multiracial or who do not identify with the categories listed. Before analysis, we decided to exclude authors who were based outside the United States ( $n = 33,098$ ), who identified as ‘other’ or who did not report ethnicity or race.

Of our sample of US-based authors who reported their race and ethnicity, more than 99% ( $n = 38,716$ ) identified as female or male (the third option was ‘prefer not to answer’). We appreciate that this binary treatment does not incorporate the full spectrum of gender and sexual identity.

### Under-represented groups

Minority ethnic and racial groups make up 31% of the US population<sup>8</sup>. People from these groups are under-represented in the STEM workforce (11%), and specifically in the physical

sciences, at 9% (ref. 9). In the AGU abstracts data set, African American, Hispanic/Latino, Native American and Pacific Islander comprise 7.7% of the first-author abstracts in the analysed sample. We combined them into one measure – under-represented minority groups (URMs). We did so to increase the statistical power to detect differences, to limit the risk of multiple comparisons generating false positives and to avoid including potentially identifying information for people from rare groups. We admit that this approach erases meaningful differences in lived experiences between people in these groups, particularly those with the lowest representation. Scientists from each minority group have unique barriers to participation.

We combined the groups White and Asian American into non-URM. We did so because Asian Americans (4.8% of the US population<sup>8</sup>) are well represented in the STEM workforce (20.6%), in physical sciences (17.5%)<sup>9</sup> and in the analysed sample (10.2% of first-author abstracts). We appreciate that this bracketing, too, is suboptimal – it also erases many meaningful differences, pressingly that Asian American researchers do face career barriers, including implicit and explicit biases<sup>10,11</sup> (see Supplementary information).

### Results

Our analyses focus on the chances of scientists from minority racial and ethnic groups that are under-represented in Earth and space sciences being given speaking opportunities, compared with other applicants. The key proportions are normalized relative to the population of each group, so that the results indicate representation (see Supplementary information for all inferential statistics).

First authors from under-represented minority groups contributed 7.7% of all the abstracts in the sample ( $n = 2,981$ ; see ‘Fewer abstracts’). The URM applicants were disproportionately students or early-career scientists (79% compared with 59% of non-URM authors; see ‘Fewer seniors’). At some career stages, the small number of URM researchers sometimes led to low statistical power to detect differences.

URM authors were invited to give talks less often than were other authors (8% versus 14%, normalized; see ‘Too few talks’). Crucially, this was statistically significant in the early-career stage (and overall).

From talk-or-poster submissions, URM authors were assigned talks less frequently than were other scientists (42.9% versus 50.8% normalized in each population; see ‘Too few

## Comment

talks'). Again, this difference was statistically significant in the crucial early-career stage as well as overall. Compared to others, URM authors were more likely to apply to give only a poster (35% versus 24%; see 'Too few talks'). This was significant overall and for each career stage.

Female URM authors had strikingly few opportunities at the AGU Fall Meetings. They had even less chance of being invited to talk (and applied for posters more often) than had URM men (and non-URM women), and were assigned talks less often than were non-URM women (see 'Fewest chances'). This is despite the fact that women (taking all races and ethnicities together) had equal or more opportunities to speak than men had (see 'Equity – why so slow?')<sup>5</sup>.

To sum up, scientists from under-represented racial and ethnic minority groups had the smallest chances of being selected and invited to speak, and opted for poster presentations more often than did their peers.

### Caveats and confounders

We did not assess abstract quality. An alternative explanation for our results could be that URM scientists submitted abstracts of lower quality. Even if the AGU's selection were perfectly meritocratic, any gap in abstract quality would still, in our view, suggest bias in the STEM pipeline – for example, as a result of discrimination in earlier education<sup>7</sup> and career development. These obstacles result in fewer URM scientists than scientists from other groups holding positions at elite institutions that provide excellent resources and strong collaborators.

Because of small sample sizes, it was not possible to control for career stage when we analysed by gender (see 'Fewest chances').

We did not investigate why URM geoscientists applied to give only a poster more often than did others overall, and at every career stage. There could be several reasons. People might be held back by psychological factors such as lower self-confidence<sup>12</sup>. For example, people from under-represented minority groups often report 'impostor syndrome' – feeling isolated and vulnerable in academia because they perceive themselves as having lower competence than their peers<sup>11</sup>. Or, some URM scientists might value poster presentations – this format could align with different goals, interests or lived experiences, for example by enabling researchers to communicate findings in one-on-one conversations.

Because we left out of our analysis those based outside the United States, those who identified as 'other' and those who did not report ethnicity or race, our results will probably have excluded relevant individuals – people who identify as multiracial, for example. Our main analyses therefore represent a conservative test of speaking opportunities

between minority and majority groups.

Notably, combining Asian Americans with under-represented minority groups would have yielded figures that, at face value, looked more representative. We did not do this because the US National Science Foundation (NSF) does not include Asian Americans as an under-represented group in STEM; its policy efforts are focused on

the under-represented minorities we track here. In the Supplementary information, we report separate exploratory analyses concerning Asian Americans, and examine career stage further, because of geoscience-specific nuances in the recruitment and representation of people who identify thus<sup>10</sup>.

We must also point out that other nations might apply different census definitions to

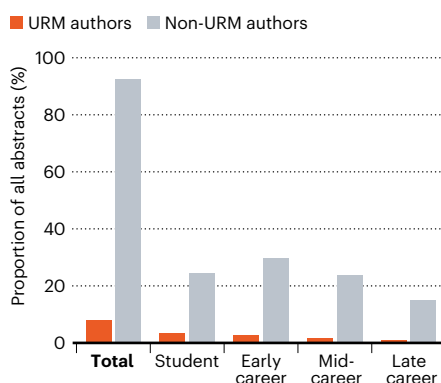
## WHO GETS THE MICROPHONE?

Some minority scientists who are already under-represented in science, technology, engineering and mathematics (STEM) and in geoscience are increasingly under-represented at every step on the path to speaking at the American Geophysical Union's Fall Meeting – in terms of abstract submissions, seniority and being offered talks.

### SUBMISSIONS

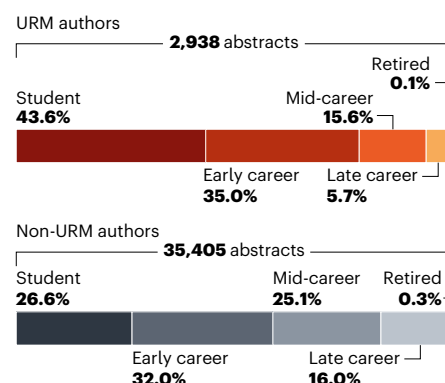
#### Fewer abstracts

Authors from under-represented minority groups (URMs; see main text for definitions) submitted the smallest proportion of abstracts in total and by career stage.



#### Fewer seniors

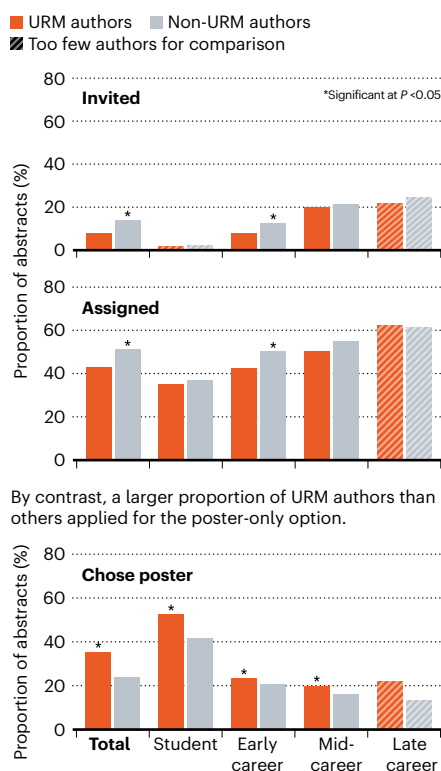
Among URM authors, a bigger proportion of abstracts are submitted by students and early-career scientists than from non-URM authors at these career stages.



### OPPORTUNITIES

#### Too few talks

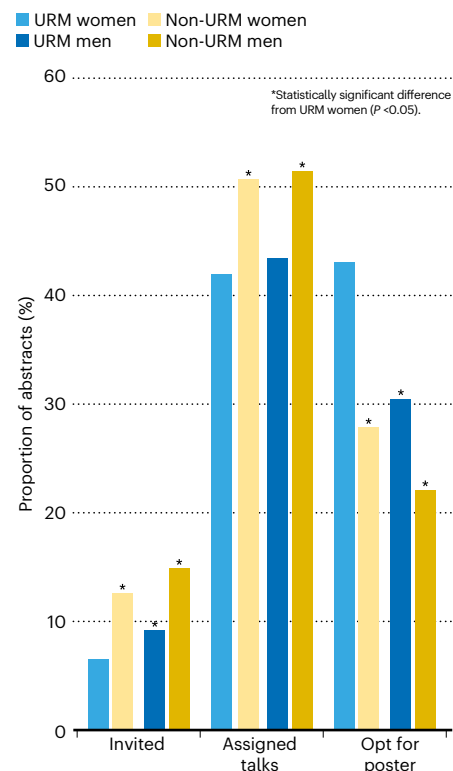
URM authors were invited or assigned to speak less often than were other authors, at most career stages.



By contrast, a larger proportion of URM authors than others applied for the poster-only option.

### FEWEST CHANCES

URM women comprise the group that is least likely to be invited or assigned to speak. But they are over-represented in requesting to present posters.



SOURCE: AM. GEOPHYS. UNION (DATA); H. L. FORD, C. BRICK ET AL. (ANALYSIS)



those used here. For example, 'White' in the United States encompasses people who have origins in the Middle East or North Africa.

## Next steps

To recap: a woman starting out in her career from a racial or ethnic minority group that is under-represented in US geoscience is less likely to gain a speaking slot at the field's largest conference than are her male peers and her non-Hispanic White peers of both sexes. These findings hold sobering lessons for the AGU and other STEM conferences and activities. We pre-registered our data cleaning and main confirmatory analyses at the Open Science Framework to increase generalizability (see Supplementary information).

One of the AGU's goals for inviting speakers is to "enhance diversity and/or feature early-career scientists". It is particularly concerning that where URM authors are most numerous – in the least-established career stages – they get fewer invitations than their proportion would predict. Such early inequities are likely to affect the retention and promotion of people from under-represented minority groups across geoscience.

There are three clear steps for the AGU to take. First, conference conveners should be blinded to information that is not necessary to rate the quality of submissions. Identifying details such as names and institutions introduce bias<sup>13,14</sup> even in people committed to equity, because many thinking processes, such as stereotype activation, occur outside awareness or control. Double-blind review has decreased bias in allocating time on the Hubble Space Telescope<sup>15</sup>.

Second, the AGU should encourage more scholars from under-represented minority groups to participate as conveners. Third, the AGU should provide more travel grants to URM presenters, which could increase the overall population of URM attendees both directly and by shifting norms. We encourage other STEM conferences to make these changes.

Meanwhile, the rest of the community has work to do to (see 'Equity – why so slow?'). Established scholars can support scientists from minority groups by encouraging them to submit talk abstracts and by providing opportunities to practise presenting in local, domestic and international venues. These steps can increase confidence and foster the development of people's identity as scientists.

It is crucial for universities and funding agencies to support organizations that provide openings and mentorship to young scholars from minority groups, such as the Society for Advancement of Chicanos/Hispanics and Native Americans in Science. The NSF aims to broaden participation in STEM through its criteria for grant proposals and through initiatives such as NSF INCLUDES (Inclusion across the Nation of Communities of Learners

## Equity – why so slow?

**Laws, policies, training, research and tracking must benefit all.**

In the United States, affirmative action is a set of laws, guidelines and policies that aim to increase the representation of historically excluded groups in higher education and professional careers. Overall, White women have been the primary beneficiaries<sup>17</sup>, as our results underscore.

A report last year by the US National Science Foundation showed that minority ethnic and racial groups are under-represented in graduate programmes, and that this results in reduced economic and social opportunities<sup>16</sup>.

An inclusive environment, visible role models and adequate funding are key to enabling people from under-represented minority groups to participate and succeed in science, technology, engineering and mathematics (STEM)<sup>18</sup>. A growing body of research has highlighted the subtle, indirect and often unintentional actions perpetrated against such researchers by majority groups, and which have an impact on a sense of belonging in STEM spaces<sup>19–21</sup>, as well as on career persistence and well-being<sup>22,23</sup>.

Small interventions can help, such as asking STEM community members to be mindful of equity, diversity and inclusion. Reminding individuals, particularly men, to consider diversity when selecting potential reviewers can improve gender representation<sup>24</sup>.

However, the effects of these reminders on ethnicity bias have not been studied, and reminders might not be effective in the long term in reducing implicit biases in STEM<sup>25</sup>. Implicit-bias training is well-meaning but largely ineffective<sup>26,27</sup>. **H.L.F., C.B. et al.**

of Underrepresented Discoverers in Engineering and Science)<sup>16</sup>. Such programmes can liaise with professional societies.

Racial, ethnic and gender biases harm individuals and undermine the quality of science. Even if all demographic gaps were plugged tomorrow at the level of people graduating with PhDs, and even if these graduates did not have to run the gauntlet of systematic bias that their predecessors faced, it could still take generations to achieve fair representation among senior academics.

We therefore urge more organizations to measure and share the outcomes for scholars from minority groups. With this information

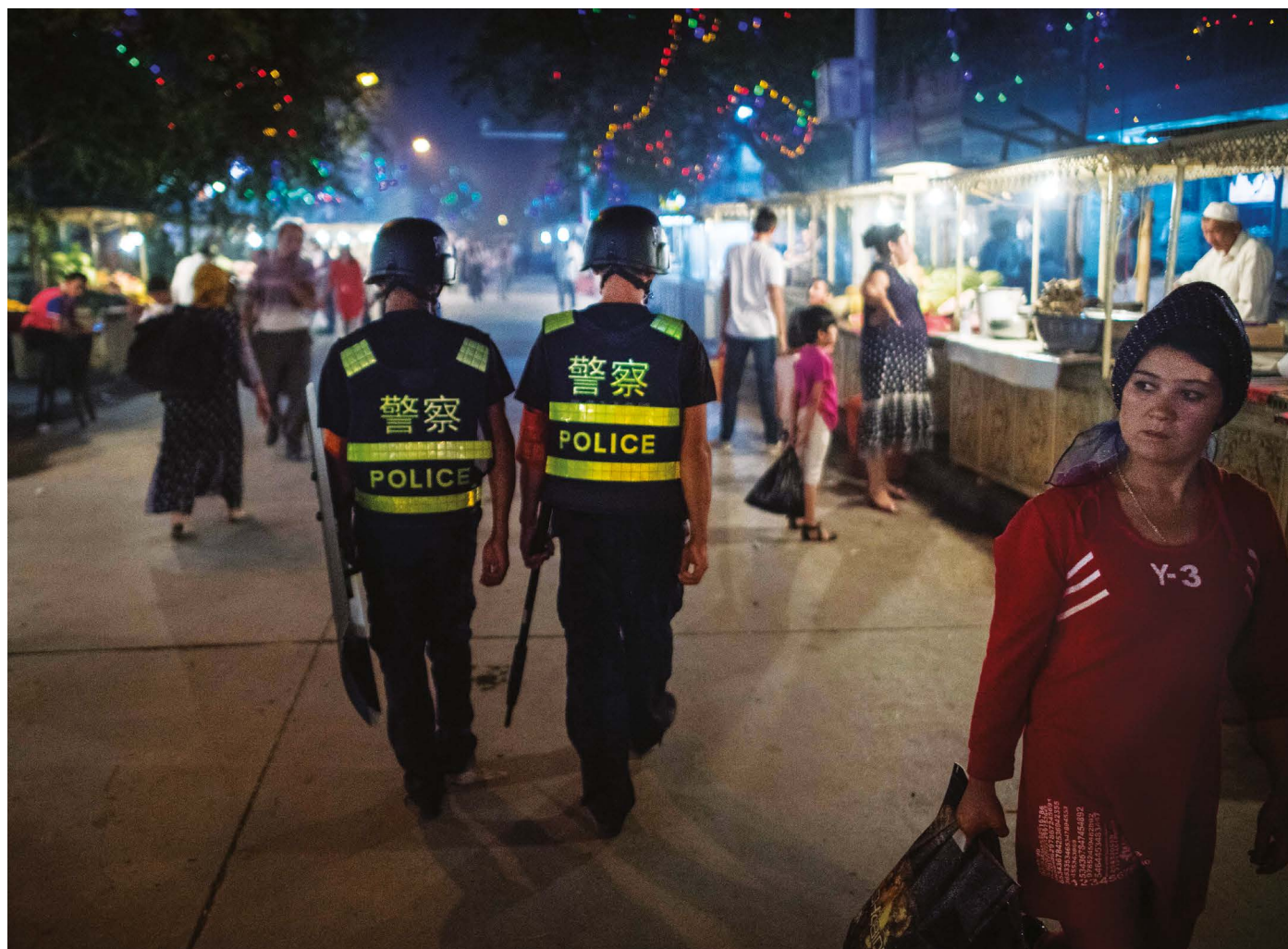
and the growing literature on effective interventions, together we can create a more equitable scientific community.

## The authors

**Heather L. Ford** is a lecturer at the School of Geography, Queen Mary University of London, UK. **Cameron Brick** is a research associate in the Department of Psychology, University of Cambridge, UK, and an assistant professor of social psychology in the Department of Psychology, University of Amsterdam, the Netherlands. **Margarita Azmitia** is a professor in the Psychology Department, University of California Santa Cruz, USA. **Karine Blaufuss** is director of business intelligence and data at the American Geophysical Union, Washington DC, USA. **Petra Dekens** is a professor in the Department of Earth & Climate Sciences, San Francisco State University, California, USA. e-mails: h.ford@qmul.c.uk; c.brick@uva.nl

1. US National Science Foundation. *Women, Minorities, and Persons with Disabilities in Science and Engineering: 2017*. Special Report NSF 17-310 (NSF, 2017).
2. Stewart, A. J. & Valian, V. *An Inclusive Academy: Achieving Diversity and Excellence* (MIT Press, 2018).
3. Crenshaw, K. W. *Stanford Law Rev.* **43**, 1241–1299 (1991).
4. Kalejta, R. F. & Palmenberg, A. C. *J. Virol.* **91**, e00739–17 (2017).
5. Ford, H. L., Brick, C., Blaufuss, K. & Dekens, P. S. *Nature Commun.* **9**, 1358 (2018).
6. Nitttrouer, C. L. et al. *Proc. Natl Acad. Sci. USA* **115**, 104–108 (2018).
7. Milner, H. R. IV *J. Black Stud.* **43**, 693–718 (2012).
8. Humes, K. R., Jones, N. A. & Ramirez, R. R. *Overview of Race and Hispanic Origin: 2010* (US Census Bureau, 2011).
9. US National Science Board. *Science and Engineering Labor Force* (NSF, 2018).
10. Ibarra, D. E., Lau, K. V., Bernard, R. E. & Cooperdock, E. H. G. Preprint at Earth and Space Science Open Archive <https://doi.org/10.1002/essoar.10500088.1> (2018).
11. Dancy, T. E. & Brown, M. C. *J. Sch. Leader.* **21**, 607–634 (2011).
12. MacPhee, D., Farro, S. & Canetto, S. S. *Anal. Soc. Issues Public Pol.* **13**, 347–369 (2013).
13. Hall, W. J. et al. *Am. J. Public Health* **105**, e60–76 (2015).
14. Axt, J. & Lai, C. K. *J. Personal. Soc. Psychol.* **117**, 26–49 (2019).
15. Strolger, L. & Natarajan, P. *Phys. Today* <https://doi.org/10.1063/PT.6.3.20190301a> (2019).
16. US National Science Foundation. *NSF INCLUDES: Report to the Nation NSF 18-040* (NSF, 2018).
17. Crenshaw, K. W. *Mich. Law Rev. First Impr.* **105**, 123–133 (2007).
18. Syed, M., Azmitia, M. & Cooper, C. R. *J. Soc. Issues* **67**, 442–468 (2011).
19. Grossman, J. M. & Porche, M. V. *Urban Educ.* **49**, 698–727 (2014).
20. Burt, B. A., Mcken, A. S., Burkhart, J. A., Hormell, J. & Knight, A. J. *ASCE Peer* <https://doi.org/10.18260/p.26029> (2016).
21. Rattan, A. & Dweck, C. S. *J. Appl. Psychol.* **103**, 676–687 (2018).
22. Ong, M., Smith, J. M. & Ko, L. T. *J. Res. Sci. Teach.* **55**, 206–245 (2018).
23. Wilkins-Yel, K. G., Hyman, J. & Zounlome, N. O. O. *J. Vocational Behav.* **113**, 51–61 (2018).
24. Hanson, B. & Lerback, J. *Eos* <https://doi.org/10.1029/2017EO083837> (2017).
25. Lai, C. K. et al. *J. Exp. Psychol. Gen.* **145**, 1001–1016 (2016).
26. Duguid, M. M. & Thomas-Hunt, M. C. *J. Appl. Psychol.* **100**, 343–359 (2015).
27. Dobbin, F., Schrage, D. & Kalev, A. *Am. Sociol. Rev.* **80**, 1014–1044 (2015).

Supplementary information accompanies this article: see [go.nature.com/2sqkqjaf](https://go.nature.com/2sqkqjaf)



JOHANNES EISELE/AFP/GETTY

Police patrol a food market at night in Kashgar in China's Xinjiang province.

# Crack down on genomic surveillance

Yves Moreau

Corporations selling DNA-profiling technology are aiding human-rights abuses. Governments, legislators, researchers, reviewers and publishers must act.

**A**cross the world, DNA databases that could be used for state-level surveillance are steadily growing.

The most striking case is in China. Here police are using a national DNA database along with other kinds of surveillance data, such as from video cameras and facial scanners, to monitor the minority Muslim Uyghur population in the western province of Xinjiang.

Concerns about the potential downsides of governments being able to interrogate people's DNA have been voiced since the early 2000s (ref. 1) by activist groups, such as the non-profit organization GeneWatch UK, and some geneticists (myself included). Partly thanks to such

debate, legislation and best practices have emerged in many countries around the use of DNA profiling in law enforcement<sup>2</sup>. (In profiling, several regions across the genome, each consisting of tens of nucleotides, are sequenced to identify a person or their relatives.)

Now the stakes are higher for two reasons. First, as technology gets cheaper, many countries might want to build massive DNA databases. Second, DNA-profiling technology can be used in conjunction with other tools for biometric identification – and alongside the analysis of many other types of personal data, including an individual's posting behaviour on social networks. Last year, the Chinese firm Forensic Genomics International (FGI) announced that it was storing the DNA profiles of more than 100,000 people from across China (FGI, known as Shenzhen Huada Forensic Technology in China, is a subsidiary of the BGI, the world's largest genome-research organization). It made the information available to the individuals through WeChat, China's equivalent of WhatsApp, using an app accessed by facial recognition.

With stringent safeguards and oversight, it is legitimate for law-enforcement agencies to



use DNA-profiling technology. But these uses can easily creep towards human-rights abuses. In October this year, the US Department of Homeland Security announced that it would authorize the mandatory collection of DNA samples from immigrants in federal custody at the US border, including children and those applying for asylum at legal ports of entry. The resulting DNA profiles will be available through a database called CODIS (Combined DNA Index System), which includes the profiles of convicted offenders and individuals arrested for serious offences. Such treatment could reinforce debunked claims that immigrants are more prone to criminal behaviour than the general population.

A much broader array of stakeholders must engage with the problems that DNA databases present. In particular, governments, policymakers and legislators should tighten regulation and reduce the likelihood of corporations aiding potential human-rights abuses by selling DNA-profiling technology to bad actors – knowingly or negligently. Researchers working on biometric identification technologies should consider more deeply how their inventions could be used. And editors, reviewers and publishers must do more to ensure that published research on biometric identification has been done in an ethical way.

### Government monitoring

In Xinjiang in China, police collected biometric information (including blood samples, fingerprints and eye scans) from nearly 19 million people in 2017, in a programme called ‘Physicals for All’. This was part of a suite of measures that are being used by the Chinese government to control the Uyghur ethnic group<sup>3</sup>.

Other nations are building massive DNA databases or considering doing so. In 2015, Kuwait passed a law mandating DNA profiling of its entire population. Foreigners living in Kuwait and even visitors were to be included. In January this year, Kenya passed a law that would have enabled the government to require all citizens to submit any biometric information, including DNA profiles, to a national database.

Both cases have hit obstacles. Kuwait’s Constitutional Court overruled the 2015 law two years later, because of concerns about how the database could be used in violations of privacy and due process. And, thanks to a decision taken by Kenya’s High Court in April, DNA is now excluded from national efforts to collect biometric data.

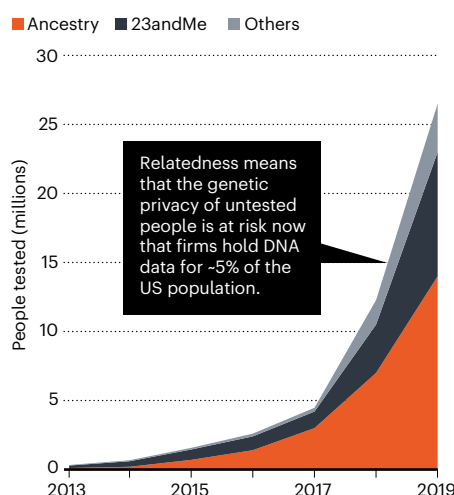
But these and other examples indicate that governments keep being tempted to Hoover up their citizens’ DNA data<sup>4</sup>.

### Corporate responsibility

One way to reduce the likelihood of massive DNA databases being misused is to change the behaviour of the companies that invest

### DNA TESTING FOR ALL

An increasing number of people are having their DNA analysed by consumer-genomics companies.



in DNA-profiling technologies (see ‘Ethical divesting’).

US and European corporations are still the dominant providers of such technologies. The deployment of DNA-surveillance infrastructure in Xinjiang, for example, was enabled by the Chinese government buying products from – and working with – the US company Thermo Fisher Scientific in Waltham, Massachusetts. The firm is currently the global leading supplier of DNA-profiling technology in law enforcement. Thermo Fisher Scientific researchers have worked with China’s Ministry of Justice, and with researchers at the People’s Public Security University of China, which falls directly under the Ministry of Public Security, to tailor the technology specifically for

### “Governments keep being tempted to Hoover up their citizens’ DNA.”

use in Tibetan and Uyghur populations<sup>5</sup>. (Thermo Fisher Scientific did not respond to a request for comment). However, in February, after two years of public outcry and intense pressure from high-profile US senators, the company announced that it would stop selling its DNA-profiling technology in Xinjiang.

Marketing and lobbying by technology suppliers is often behind pushes for the broadest possible use of DNA profiling. In 2016, for instance, a representative of a US lobbying firm working for Thermo Fisher Scientific described in a conference presentation the development of universal DNA databases as “inevitable”. He noted that the expansion of these to “Western countries or other countries with democratic forms of government” faced “significant hurdles”, such as the “open and public parliamentary process” and the

“culture of being influenced by opposition and protests” (see [go.nature.com/337pjce](https://go.nature.com/337pjce)).

Restrictions on the use of technologies or services provided by corporations are currently too weak. Take export controls: either they do not pay due attention to these sensitive technologies, or they have loopholes that often render them useless. For example, US laws forbid the export of fingerprint-recognition technology to some destinations or users deemed problematic by the US government, such as the Chinese police. But the United States does not restrict the export of more-invasive DNA-profiling and facial-recognition technologies. Meanwhile, the European Union does not regulate the export of fingerprint technology, even though the dominant global suppliers are European.

Export controls for biometric technologies could be improved relatively easily. The US Department of Commerce is currently considering revising regulations for emerging technologies<sup>6</sup>, such as Internet censorship and video surveillance, to try to reduce the likelihood of companies doing business with problematic buyers. Last month, it barred Xinjiang police forces and eight Chinese technology companies from buying US products or importing US technology because of their role in the repression of Uyghurs.

Some regulatory initiatives are promising and could provide a deterrent if enforced. The 2017 EU directive on non-financial reporting (named 2014/95) has mandated that large companies listed on stock markets document their social and environmental impacts in their annual reports for shareholders and the public. Since 2017, France’s corporate ‘duty of vigilance’ law has required all French companies employing more than 5,000 people in the country to actively monitor their impacts on human rights, the environment and so on (see [go.nature.com/208tcvn](https://go.nature.com/208tcvn)).

In the United States, several human-rights lawyers have attempted to revive the Alien Tort Statute (28 U.S.C. § 1350) over the past 20 years. Produced in 1789 but never deployed, this law could enable a foreign individual to make a civil liability claim against a domestic corporation in US courts. A carefully crafted Alien Tort Statute could provide a way to hold companies to the same standards, whether they are operating at home or abroad.

Ultimately, international laws must be established that clearly stipulate the human-rights responsibilities of corporations. For the past decade, a United Nations working group has been drafting a treaty to regulate the activities of transnational corporations with regards to human rights and the environment (see [go.nature.com/35qnehe](https://go.nature.com/35qnehe)). If it is not crippled by lobbying, this could eventually become a powerful tool to promote

ethical business practices. Yet companies are only part of the story when it comes to the potential misuse of DNA databases.

## Research ethics

The chain of technology development leads from fundamental to applied research to the products that enable the abuses. More academics working on biometric identification technology should reflect on the potential misuses of their inventions and engage with society. For instance they can contribute to mainstream media, participate in public debates or join ethics boards.

Recent events indicate that publishers and scholars might be paying insufficient attention to the sources of biometric-identification research. For example, in August last year, after several Human Rights Watch and media reports about the surveillance abuses in Xinjiang, Springer Nature published the proceedings of a biometrics conference held in the province. (Springer Nature has been the publisher of the proceedings of the Chinese Conference on Biometric Recognition for nine years; *Nature* is editorially independent of its publisher.) One of the conference papers, on technologies for recognizing various languages in images, described how “Uyghur information” (referring to the Uyghur language script) could be detected in images that might be used to evade Internet censorship<sup>7</sup>. Another paper described how products from Thermo Fisher Scientific and the Chinese firms Hisign, Megvii and iFlytek are being used to build a population-scale database for DNA, fingerprint, face and voice information in a major Chinese city<sup>8</sup>.

In July this year, researchers from Imperial College London announced the results of an open competition on facial recognition. (The winners presented their work at a conference in Seoul in October.) Before a reporter from the non-profit news platform Coda pointed it out, one of the sponsors of the conference had been a Chinese artificial-intelligence start-up called DeepGlint, which in 2018 set up a joint research laboratory with the Xinjiang police. The conference organizers removed DeepGlint as a sponsor in August.

Over the past eight years, three leading forensic genetics journals — *International Journal of Legal Medicine* (published by Springer Nature), and *Forensic Science International and Forensic Science International: Genetics Supplement Series* (both published by Elsevier) — have published 40 articles co-authored by members of the Chinese police that describe the DNA profiling of Tibetans and Muslim minorities, including people from Xinjiang. I analysed 529 articles on forensic population genetics in Chinese populations, published between 2011 and 2018 in these journals and others. By my count, Uyghurs and Tibetans are 30–40 times more frequently studied than are people from

## ETHICAL DIVESTING

### Investors could help to ensure ethical use of the products of DNA profiling firms.

Public outcry can lead to divestment. Since March this year, for example, major US funds such as Goldman Sachs have divested all their shares from the Chinese surveillance company Hikvision, because of concerns about the use of the company's products in human-rights breaches.

Investors could even be motivated to scrutinize company ethics, thanks to studies over the past five years or so indicating that ‘good’ corporate social responsibility practices tend to correlate with better financial performance over the long term.

Pressure from investors — and the public in general — might be increasingly powerful. Take Thermo Fisher Scientific's February announcement that it would stop selling its DNA profiling technology in Xinjiang, China. Although Chinese authorities can easily transport such technology from elsewhere in the country, it is significant that a major corporation publicly acknowledged “the importance of considering how [its] products and services are used — or may be used — by [its] customers”. **Y.M.**

Han communities, relative to the size of their populations (unpublished data). Half of the studies in my analysis had authors from the police force, military or judiciary. The involvement of such interests should raise red flags to reviewers and editors.

In short, the scientific community in general — and publishers in particular — need to unequivocally affirm that the Declaration of Helsinki (a set of ethical principles regarding human experimentation, developed for the medical community) applies to all biometric identification research (see [go.nature.com/34bypbf](http://go.nature.com/34bypbf)). Unethical work that has been published in this terrain must be retracted.

## Privacy concerns

DNA databases in local police forces are proliferating, even in countries that have democratic governments and well-established legal protections for citizens' privacy<sup>9</sup>. By August this year, for instance, the Office of the Chief Medical Examiner of New York City held more than 82,000 genetic profiles. At the same time, there has been a growth in consumer and recreational genomic services, such as the US corporations 23andMe in Mountain View, California, and Ancestry in Lehi, Utah (see ‘DNA testing for all’). Medical DNA sequencing is also becoming routine<sup>10</sup>.

Currently, only some consumer-genomics companies have willingly shared people's DNA data with law-enforcement agencies. And in many countries, patients' data are confidential.

But to deploy DNA surveillance across a group of people, you need profiles from only 2–5% of that population, because biological relationships can be inferred<sup>11,12</sup>. And as genealogy and medical databases mushroom, law enforcers and others are increasingly tempted to tap into them<sup>13</sup>. In 2017 in the Netherlands, the Ministry of Health drafted a bill that would have allowed police to obtain people's DNA information from hospitals in some limited cases. It was abandoned following public outcry.

And June saw what might be a game changer in the United States. The Orlando Police Department obtained a warrant that allowed it to search the entire DNA database of the GEDMatch genealogy website, based in Lake Worth, Florida. Because consumer-genomics companies already hold DNA data for an estimated 5% of the US population, unfettered access to these data by law-enforcement agencies would simply spell the end of genetic privacy in the United States.

All of us must beware a world in which our behavioural, financial and biometric data, including our DNA profiles, or even entire genome sequences, are available to corporations — and so potentially to law enforcers and political parties. Without the changes outlined here, the use of DNA for state-level surveillance could become the norm in many countries.

## The author

**Yves Moreau** is a computational biologist specializing in human genetics and professor of engineering at the Catholic University of Leuven (KU Leuven), Leuven, Belgium. e-mail: [yves.moreau@kuleuven.be](mailto:yves.moreau@kuleuven.be)

- Wallace, H. M., Jackson, A. R., Gruber, J. & Thibedeau, A. D. *Egypt. J. Forensic Sci.* **4**, 57–63 (2014).
- Forensic Genetics Policy Initiative. *Establishing Best Practice for Forensic DNA Databases* (Forensic Genetics Policy Initiative, 2017); available at <http://dnapolicyinitiative.org/report>
- Ramzy, A. & Buckley, C. *The New York Times* (16 November 2019).
- Nelkin, D. & Andrews, L. *Sociol. Health Illn.* **21**, 689–706 (1999).
- Wang, Z. *et al. Sci. Rep.* **6**, 31075 (2016).
- Bureau of Industry and Security. *Fed. Regist.* **83**, 58201–58202 (2018).
- Aizezi, Y., Jiamali, A., Abdurixiti, R. & Ubul, K. in *Biometric Recognition. CCBP 2018* (eds Zhou, J. *et al.*). *Lecture Notes in Computer Science* **10996**, 709–718 (Springer, 2018).
- Zhu, W. J., Zhuang, C. Z., Liu, J. W. & Huang, M. in *Biometric Recognition. CCBP 2018* (eds Zhou, J. *et al.*). *Lecture Notes in Computer Science* **10996**, 198–205 (Springer, 2018).
- Mercer, S. & Gabel, J. D. *N.Y.U. Ann. Surv. Am. L.* **69**, 639–698 (2014).
- Ratner, M. *Nature Biotechnol.* **36**, 484 (2018).
- Erlich, Y., Shor, T., Pe'er, I. & Carmi, S. *Science* **362**, 690–694 (2018).
- Guest, C. *Am. U. L. Rev.* **68**, 1015–1052 (2019).
- O'Doherty, K. C. *et al. BMC Med. Ethics* **17**, 54 (2016).



# Correspondence

## Joint statement on EPA proposed rule and public availability of data (2019)

Eighteen months after articulating our concerns (J. Berg *et al. Nature* <http://doi.org/crq8>; 2018) regarding the 2018 'Strengthening Transparency in Regulatory Science' rule proposed by the US Environmental Protection Agency (EPA; [go.nature.com/2kmt7g](http://go.nature.com/2kmt7g)), we have become more concerned in response to recent media coverage and a 13 November hearing on the role of science in decision-making at the EPA. These events suggest that the proposed rule is now moving towards implementation; whether it includes amendments sufficient to address the concerns raised by us and many others remains a question.

Our previous statement on the proposed rule, authored and published by the editors-in-chief of five major scientific journals in May 2018, reflected alarm that the proposal's push for 'transparency' would be used as a mechanism for suppressing the use of relevant scientific evidence in policy-making, including public-health regulations. After the public comment period for the proposed rule closed, the EPA reported more than 590,000 comments from individuals and scientific, medical and legal groups, many of which articulated similar concerns (see [go.nature.com/2jfxhnn](http://go.nature.com/2jfxhnn)).

As leaders of peer-reviewed journals, we support open sharing of research data, but we also recognize the validity of scientific studies that, for confidentiality reasons, cannot indiscriminately share absolutely all data.

Data sets featuring personal identifiers – including studies evaluating genomes of thousands of people to characterize medically relevant genetic variants – are but one example. Such data may be critical to developing new drugs or diagnostic tools, but cannot be shared openly; even anonymized personal data can be subject to re-identification, and it has been a long-standing practice for agencies and journals to acknowledge the value of data-privacy adjustments. The principles of careful data management, as they inform medicine, are just as applicable to data regarding environmental influences on public health. Discounting evidence from the decision-making process on the basis that some data are confidential runs counter to the EPA stated mission "to reduce environmental risks ... based on the best available scientific information" (see [go.nature.com/2kqhny](http://go.nature.com/2kqhny)).

We are also concerned about how the agency plans to consider options related to existing regulations. Even if a new standard is not applied retroactively, the standard could apply when a regulation is updated; thus, foundational science from years past – research on air quality and asthma, for example, or water quality and human health – could be deemed by the EPA to be insufficient for informing our most significant public-health issues. That would be a catastrophe.

We urge the EPA to continue to adopt an approach that ensures the data used in decision-making are the best available, which will at times require consideration of peer-reviewed scientific data, not all of which may be open to all members of the public. The most relevant science, vetted through peer review, should inform public

policy. Anything less will harm decision-making that claims to protect our health.

We hope that in the end, decisions that are made to inform the proposed EPA rule will rise above any form of politics, focusing on what's best for our communities. We encourage anyone with concerns or opinions about this issue to express their views through relevant legislative channels. Whether submitting public comments to the EPA or communicating with lawmakers in Congress, it is important to emphasize that decision-making that affects us all should be informed by nothing less than the full suite of relevant science vetted through peer review.

**H. Holden Thorp** Science family of journals, Washington DC, USA.  
[hthorp@aaas.org](mailto:hthorp@aaas.org)

**Magdalena Skipper** *Nature*, London, UK.

**Veronique Kiermer** Public Library of Science (PLOS) journals, San Francisco, California, USA.

**May Berenbaum** *Proceedings of the National Academy of Sciences*, Washington DC, USA.

**Deborah Sweet** Cell Press, Cambridge, Massachusetts, USA.

**Richard Horton** *The Lancet*, London, UK.

**Editor's note:** This statement was published online on 26 November, and simultaneously as a letter in *Science* (H. Holden Thorp *et al. Science* <https://doi.org/10.1126/science.aba3197>; 2019), which should be the primary citation. It is being disseminated by other publications represented by the signatories.

## Boost glacier monitoring

Glacier-mass changes are a reliable indicator of climate change. On behalf of the worldwide network of glacier observers, we urge parties to the United Nations Framework Convention on Climate Change to boost international cooperation in monitoring these changes, and to include the results in the Paris agreement's global stocktake.

Since 1960, glaciers have lost more than 9,000 gigatonnes of ice worldwide – the equivalent of a 20-metre-thick layer with the area of Spain. This melting alone – as distinct from that of the Greenland and Antarctic ice sheets – has raised global sea level by almost 3 centimetres, contributing 25–30% of the total rise (M. Zemp *et al. Nature* **568**, 382–386; 2019).

The present rate of melting is unprecedented. Several mountain ranges are likely to lose most of their glaciers this century. And we face the loss of almost all glaciers by 2300 (B. Marzeion *et al. Cryosph.* **6**, 1295–1322; 2012).

Glacier shrinkage will severely affect freshwater availability and increase the risk of local geohazards. Global sea-level rise will result in the displacement of millions of people in coastal regions and in the loss of life, livelihoods and cultural-heritage sites.

The systematic monitoring of glaciers has been internationally coordinated for 125 years. Continuing to do so will document progress in limiting climate change for current and future generations.

**Michael Zemp\*** World Glacier Monitoring Service, University of Zurich, Switzerland.  
[michael.zemp@geo.uzh.ch](mailto:michael.zemp@geo.uzh.ch)

\* On behalf of 38 co-signatories; see [go.nature.com/34ak25y](http://go.nature.com/34ak25y)

# News & views

## Metallurgy

# Fine-grained metals from 3D printing

Amy J. Clarke

Conventional alloys have undesirably coarse-grained microstructures when used in 3D printing. A designer alloy overcomes this problem, potentially opening the way to the widespread adoption of 3D metal printing. **See p.91**

There are many potential benefits to using additive manufacturing – also known as 3D printing – for making metal parts, rather than conventional manufacturing processes. For example, additive manufacturing is highly customizable, it can produce complex structures and it can be used for the economical production of low numbers of metal components. But to achieve the strict specifications needed for some applications, the microscopic structure of printed metal objects must be controlled. On page 91, Zhang *et al.*<sup>1</sup> describe titanium–copper alloys that produce practically useful microscopic structures during additive manufacturing, removing the need for subsequent treatment. The resulting materials exhibit promising combinations of mechanical properties, comparable to those of the ubiquitous structural alloy Ti-6Al-4V, produced using conventional and additive manufacturing processes.

In metal additive manufacturing, an alloy (in the form of powders or wires) is deposited in a layer and then melted by a rapidly moving heat source to form a solid mass; successive layers are built up to produce a 3D part. The process typically produces large temperature gradients, high solidification rates and repeated cycles of heating and cooling. A common characteristic of 3D-printed metals is coarse columnar grains that grow along specific directions of the crystal lattice that are favourably oriented with the heat flow (Fig. 1a).

Coarse columnar grains are usually undesirable because they can cause the printed material to have direction-dependent (anisotropic) mechanical properties and make it susceptible to tearing or cracking during solidification<sup>2–4</sup>. However, columnar solidification can undergo a transition to equiaxed solidification – in which the grains produced

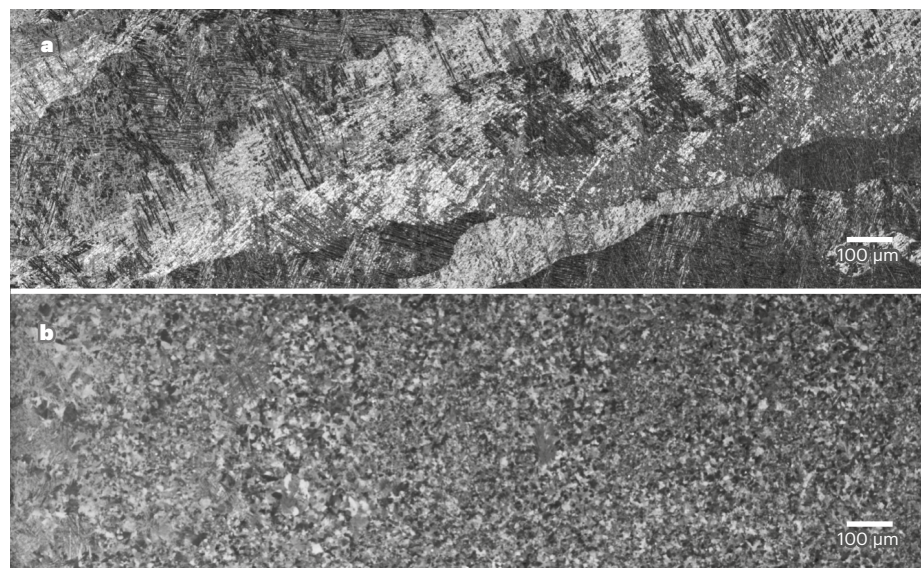
have similar dimensions in all directions – by changing the processing conditions used for additive manufacturing<sup>2</sup>. Alloys with equiaxed grains have desirably uniform properties, and so methods for producing them are of great technological value<sup>4</sup>.

Models and experiments have been used to study the columnar-to-equiaxed transition (CET) in nickel-based alloys that have been melted using an electron beam<sup>2,3</sup>. The number of nuclei (tiny crystals that ‘seed’ the growth of the solid phase) in the liquid metal, and the processing conditions used during electron-beam additive manufacturing, were found to have a larger influence on

grain structure than did the composition of the alloy<sup>3</sup>. This suggests that the CET can be controlled through process design and by promoting nucleus formation in alloy melts. Additives called inoculants, which cause nuclei to form in the melt, have been incorporated into metal–alloy powders used in additive manufacturing, to increase the density of nuclei and thereby promote the formation of equiaxed grains<sup>4</sup>. However, suitable inoculants for titanium alloys remain elusive.

Zhang *et al.* now show that fine equiaxed grains, on average less than 10 micrometres in diameter, can be produced in titanium–copper alloys during additive manufacturing, without adding inoculants (Fig. 1b). The authors propose that nucleation and CET are promoted in these alloys by the formation of a large zone of supercooled liquid – melted alloy that is fully liquid, despite its being below the temperature at which the alloy should start to solidify. The final product consists of two solid phases that contain different amounts of titanium and copper, forming a microstructure that includes nanoscale plates (lamellae). The mechanical properties of the printed material compare favourably with those of Ti-6Al-4V, and of cast (and heat-treated) titanium–copper alloys.

The authors suggest that equiaxed grains are produced during solidification of the melt, and that further microstructural refinement might then occur during the cyclical



**Figure 1 | Grain structure in printed metals.** **a**, When conventional metal alloys are used for 3D printing, large columnar grains tend to form, as shown here for the structural alloy Ti-6Al-4V. This causes the printed alloy to have undesirable anisotropic (direction-dependent) properties. **b**, Zhang *et al.*<sup>1</sup> report that titanium–copper alloys produced by 3D printing contain fine grains that have similar dimensions in all directions. The alloy shown here was produced using the same conditions as in **a**. (Images from ref. 1.)



temperature changes associated with the 3D-printing process. However, it is difficult to tell unambiguously whether the solidification step is the genesis of the fine grains, because the microstructures produced at high temperatures during solidification will be replaced by features that develop during subsequent solid-state phase transitions. Another plausible scenario is that columnar grains form during solidification, and that equiaxed grains are produced and refined during solid-state thermal cycling. Such grain refinement has been reported in steels<sup>5</sup>.

When steels that have a two-phase lamellar microstructure at low temperatures are heated above a critical temperature, new grains of a third phase (austenite) nucleate and grow. The two low-temperature phases then re-form on cooling<sup>5</sup>. Repeated nucleation and growth of the various phases can therefore occur under suitable conditions during thermal cycling, leading to significant grain refinement.

Alloys such as Ti-6Al-4V typically do not undergo grain refinement during thermal cycling<sup>6</sup>, because no new grains of the high-temperature phase nucleate. However, it is unclear whether new grains of high-temperature phase can nucleate and grow in Ti-6Al-4V during thermal cycling typical of additive manufacturing<sup>7</sup>, which might conceivably refine grains. Zhang and colleagues' titanium-copper alloys have high- and low-temperature phases analogous to those of steels. Clarifying the role of nucleation and growth of these phases in grain refinement during thermal cycling should be a topic of future research.

A deeper understanding of solidification and solid-state phase transitions is clearly needed to guide the design of future alloys for additive manufacturing and to control their microstructures – although the nucleation stage is hard to study experimentally. It is also imperative that we have a better understanding of how the rapidly changing conditions during additive manufacturing influence microstructure development. *In situ* characterization of phase transitions and dynamic phenomena, for example using imaging and diffraction techniques in experiments that simulate the conditions of additive manufacturing<sup>8,9</sup>, might help to unveil some of the complexity of the processes involved. Such efforts are timely, and are necessary to produce optimized alloys that will lead to the widespread adoption of additive manufacturing for the production of high-performance structural parts, for which reliably high-quality microstructures and mechanical properties are of the utmost importance.

**Amy J. Clarke** is in the George S. Ansell Department of Metallurgical and Materials Engineering, Colorado School of Mines, Golden, Colorado 80401, USA.  
e-mail: amyclarke@mines.edu

1. Zhang, D. *et al.* *Nature* **576**, 91–95 (2019).
2. Dehoff, R. R. *et al.* *Mater. Sci. Technol.* **31**, 931–938 (2015).
3. Haines, M., Plotkowski, A., Frederick, C. L., Schwalbach, E. L. & Babu, S. S. *Comput. Mater. Sci.* **155**, 340–349 (2018).
4. Martin, J. H. *et al.* *Nature* **549**, 365–369 (2019).
5. Karlsson, B. *Mater. Sci. Eng.* **11**, 185–193 (1973).
6. Ivasishin, O. M. & Teliovich, R. V. *Mater. Sci. Eng. A* **263**, 142–154 (1999).
7. Zhong, H. Z., Qian, M., Hou, W., Zhang, X. Y. & Gu, J. F. *Mater. Lett.* **216**, 50–53 (2018).
8. Zhao, C. *et al.* *Sci. Rep.* **7**, 3602 (2017).
9. McKeown, J. T. *et al.* *JOM* **68**, 985–999 (2016).

## Neuroscience

# The fruit fly gets oriented

Malcolm G. Campbell & Lisa M. Giocomo

Two studies in flies reveal the mechanism by which the brain's directional system learns to align information about self-orientation with environmental landmarks – a process crucial for accurate navigation. **See p.121 & p.126**

As everyone knows, a good sense of direction is needed to successfully navigate the world. In mammals, this 'sense' involves neurons called head-direction cells. Each such cell becomes most active when the animal faces a particular direction relative to landmarks in its environment. Together, the cells' activity indicates which direction the animal is facing in at any given moment. In 2015, it emerged that fruit flies, which are much easier than mammals to study experimentally, have strikingly similar cells, called heading neurons<sup>1</sup>. Fisher *et al.*<sup>2</sup> (page 121) and Kim *et al.*<sup>3</sup> (page 126) now build on this discovery to tackle a decades-old problem: how does this type of neuron respond to the locations of landmarks

similar landmarks have been seen before – the particular configuration of street signs at the new station must be learnt, even though you may have seen similar street signs in other places.

The neural mechanisms that underlie these abilities in flies are a beautiful example of form following function. The insects' heading neurons (also known as E-PG, or compass, neurons) are arranged in a ring (Fig. 1) that corresponds to the 360° of possible directions in which the fly can face<sup>1</sup>, sometimes called heading angles. Because of inhibition between neurons, only one heading angle can be indicated at one time, providing the fly with an unambiguous signal. Of note, rather than always aligning their activity to a cardinal direction such as north, heading neurons realign their activity arbitrarily when the fly enters a new environment. The heading neurons receive input from visual ring neurons, which are activated by visual cues at particular orientations relative to the fly, and from internal cues about self-motion.

Fisher *et al.* set out to test whether and how the connections between visual ring neurons and heading neurons change with experience, using a range of experimental techniques (many of which are possible only in fruit flies). They implemented a virtual-reality (VR) system in which the fly walked on a floating ball. An array of lights around the fly flashed on and off in concert with the animal's movements<sup>4</sup>, providing visual cues to enable the fly to orient itself. The authors then measured inputs from visual ring neurons to heading neurons as the flies explored this virtual environment. They also used genetic techniques to inhibit the activity of visual ring neurons.

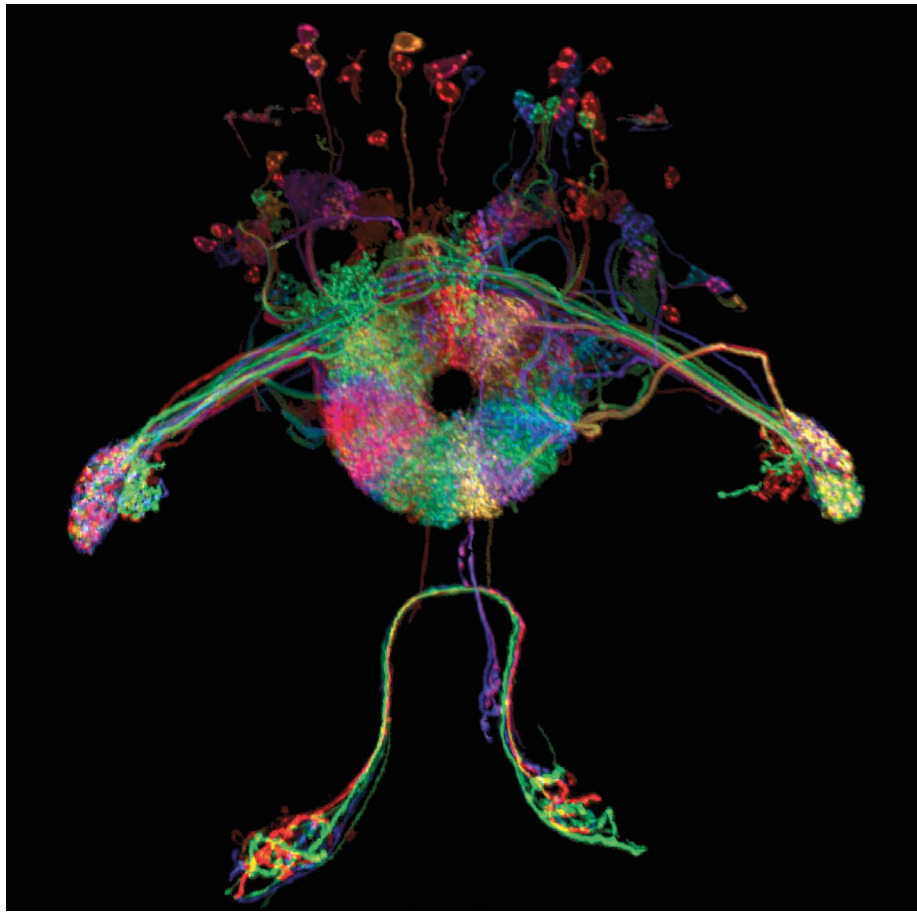
These experiments revealed that individual heading neurons are inhibited by visual ring neurons that are activated by visual cues at specific angles relative to the fly. Because of the specificity of this pairing, the visual input

**“This shows that the fly's heading network can store and retrieve memories of scenes.”**

in a manner that is stable enough to be reliable, but flexible enough to allow adaptation to new environments?

To give an example of the problem, imagine emerging from a subway station onto a crowded street. If you are a regular visitor, a glance around is all you need to be on your way. However, if you have never been to this station before, you might need a moment to orient yourself. You take note of surrounding street signs, shops and monuments. Before long, you have your bearings and can set off in the right direction.

This example highlights two challenges for the brain's directional system. First, it must stably indicate direction in familiar environments: returning to the same station should call the same orientation to mind. Second, it must have the flexibility to learn new configurations of landmarks, even when



**Figure 1 | Neurons in the central complex of the fruit-fly brain, tagged with fluorescent proteins.** The central complex includes a ring-like structure called the ellipsoid body that contains heading neurons. These cells correspond to all the possible directions in which the fly can face, providing the insect with a compass-like signal that it uses to navigate. Two studies<sup>2,3</sup> have revealed how flies orient themselves in familiar environments and adapt to new ones, thanks to signalling to heading neurons from visual ring neurons, which originate in the eyes (not shown).

reinforces the directional preference of the heading neurons. This work solves the problem of how the brain can transform visual input into a stable directional signal in a familiar environment – the first of the challenges in our subway scenario.

Next, Fisher *et al.* tested how heading neurons can adapt when their environment changes. They presented flies with two identical visual cues, separated by 180° – an ambiguous environment in which a half turn produces the same visual cue as a full turn. The flies' heading neurons, which can represent only one heading angle at a time, flipped between being preferentially activated by two opposing heading orientations.

After the flies were returned to the one-cue world, the relationship between visual input and the activity of the heading network as a whole sometimes changed by 180°. The strength of visual inputs to heading neurons also changed, but only in neurons that were active during the two-cue period.

This finding shows that new associations can form between visual ring neurons and heading neurons in new environments. However,

simple visual changes are not enough. Instead, there must be a coordinated activation of the upstream visual ring neuron and downstream heading neuron. This leads to a decrease in the strength of the inhibitory synaptic connection between them, so that the heading neuron becomes less sensitive to inhibition by the visual ring neuron – a phenomenon known as associative plasticity.

In a complementary experiment, Kim *et al.* presented flies with VR scenes derived from natural images, moving a step closer to naturalistic conditions. They then stimulated heading neurons in arbitrary orientations relative to the visual cues the fly was receiving, thereby altering neurons' preferred heading directions. After this stimulation period, the offsets between heading-neuron activity and visual input remained intact, demonstrating the capacity of the system to learn new visual–heading associations. Even partial views of a scene, when paired with stimulation, caused global changes in the activity of the heading-neuron network. This reveals a useful property of the network for our subway set-up: it enables you to orient yourself

at a new station without having to survey all 360° of the scene.

But the system's flexibility could have a downside – if synapses can change, can they also be erased? Kim *et al.* asked whether the heading network can 'remember' multiple scenes. First, they found that presenting flies with different scenes elicited different heading-neuron direction preferences, which varied from fly to fly. But, crucially, these preferences were stable for a given scene for each fly, even when the scene was presented as part of a 'slide show' of multiple different scenes. This shows that the fly's heading network can store and retrieve memories of scenes. The authors conclude their paper by developing theories that predict what types of scene can be simultaneously stored and what kinds of rule allow scenes to be learnt without existing memories being erased.

Together, these studies rigorously establish the ability of the fly's heading network to learn through associative plasticity. Future work should explore the memory capacity of the system. A key question is whether flies and other insects use memories of complex scenes for navigation, or rely more heavily on celestial cues such as the Sun<sup>5</sup>. Other types of sensory input, such as light polarization, also probably have a role in anchoring insect heading representations, and need to be taken into account. In addition, molecular and cellular work will be needed to uncover the synaptic-plasticity rules at work in the system and to determine whether they match Kim and colleagues' theoretical prediction. Finally, this work generates hypotheses that should be tested in other species, because many properties of the fruit fly's heading neurons are similar to those of mammalian head-direction cells.

So, although it might not have mastered the subway, the fruit fly has deepened our understanding of the neural mechanisms that underlie our sense of direction. A rich landscape of further research awaits.

**Malcolm G. Campbell** and **Lisa M. Giocomo**

are in the Stanford University School of Medicine, Stanford, California 94305, USA.

**M.G.C.** is also in the Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts.

e-mails: mgcampb@fas.harvard.edu;

giocomo@stanford.edu

1. Seelig, J. D. & Jayaraman, V. *Nature* **521**, 186–191 (2015).
2. Fisher, Y. E., Lu, J., D'Alessandro, I. & Wilson, R. I. *Nature* **576**, 121–125 (2019).
3. Kim, S. S., Hermundstad, A. M., Romani, S., Abbott, L. F. & Jayaraman, V. *Nature* **576**, 126–131 (2019).
4. Strauss, R., Schuster, S. & Götz, K. G. *J. Exp. Biol.* **200**, 1281–1296 (1997).
5. Wehner, R. *Annu. Rev. Entomol.* **29**, 277–298 (1984).

This article was published online on 20 November 2019.



## Evolution

# All ears about ancient mammals

Anne Weil

The configuration of middle-ear bones in an ancient fossil suggests that specializations suited to eating plants might have influenced how the jaw joint evolved to form the mammal's ear. **See p.102**

The presence of three delicate bones in the middle ear that are completely separated from the lower jaw can be used to distinguish existing mammals from other vertebrates. This arrangement evolved independently at least three times in mammals, so it is not found in all mammalian fossils. On page 102, Wang *et al.*<sup>1</sup> describe a newly discovered fossil that reveals how these different middle ears evolved into distinct configurations.

The authors named this previously unknown species *Jeholbaatar kielanae*. It was about the size of a vole, and scampered around China about 120 million years ago. It belonged to the longest-lived mammalian lineage, the multituberculates. These typically small-bodied mammals persisted from about 160 million to 34 million years ago, and diverse members of this lineage became common throughout the Northern Hemisphere<sup>2</sup>.

Multituberculates might have been so successful because they chewed differently from other mammals. Instead of slicing food into pieces using a vertical biting motion like a cat does, or grinding their food by moving their lower jaw (the mandible) horizontally and sideways like a cow, multituberculates sliced and ground food by drawing their mandible horizontally but backwards. This innovation, 'palinal motion', required specializations of the teeth, jaw joint and musculature. It contributed to the unmatched longevity of the multituberculate lineage, and it facilitated group diversification by enabling multituberculates to use plants as a food source at a time in prehistory when other mammals mainly ate insects or small vertebrates.

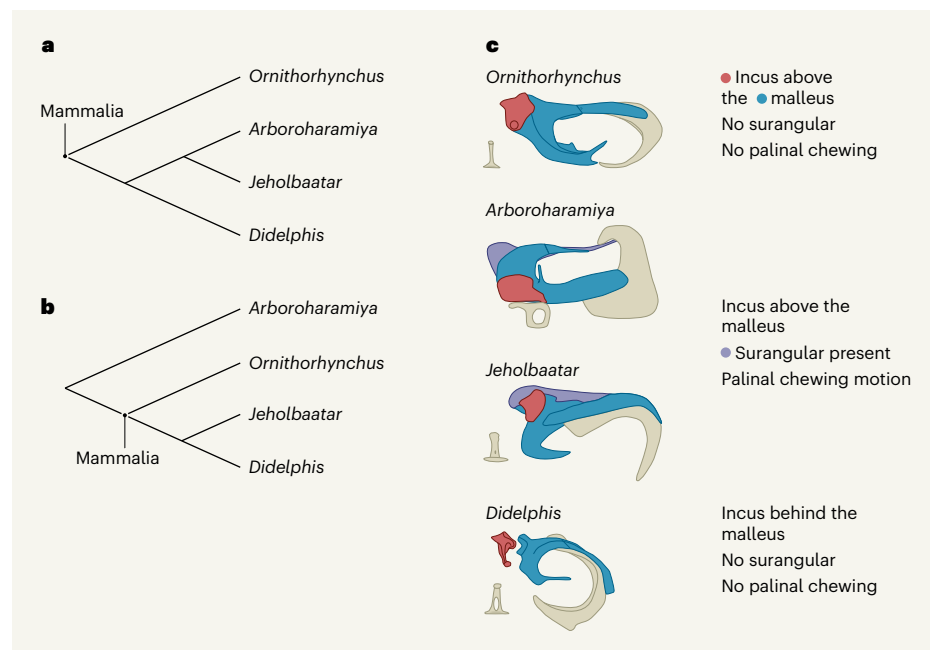
Wang and colleagues argue that the adaptation of this chewing approach also drove the evolution of an unusual type of ear. In each independent instance, mammalian middle ears evolved from an ancestral jaw joint. In every case, the articular bone at the back of the mandible and the quadrate bone (which became the incus bone of the middle ear) that it made contact with on the skull retained their connection. These bones shifted slightly

internally to form a middle ear together with a bone called the stapes, which was present in mammalian ancestors. Other bones then formed the jaw joint that mammals have today. In transitional stages of this evolutionary process, the connection between the middle ear and the mandible was still present at a middle-ear bone called the malleus, although the extent of this connection was reduced compared with the connection in the ancestral state<sup>3</sup>. Both the jaw and the ear had to function at

all stages of the transition. If multituberculates had adopted palinal chewing before the separation of the middle-ear bones from the jaw, how would this arrangement have worked? The tiny but exquisitely preserved middle ear of *Jeholbaatar* (Fig. 1) is completely separated from the jaw, but it provides the beginning of an answer to this question.

It has long been suspected that, in mammalian ancestors, the articular bone and the prearticular bone of the ancestral jaw fused to form the malleus. Fossil discoveries have suggested that a third bone, the surangular, also fused with the articular, at least in some lineages<sup>3,4</sup>. In *Jeholbaatar*, the surangular is present as a separate bone distinguishable along the lateral side of the malleus. The only other animal in which a separate surangular has been described in the ear also shares a second odd trait with *Jeholbaatar*<sup>4</sup>: the position of the incus in the middle ear.

The incus lies flat on top of the malleus in *Jeholbaatar*, in contrast to its position in humans and opossums (*Didelphis*), in which it is positioned posteriorly, behind the malleus. This contact between the incus and the malleus in *Jeholbaatar*, horizontal and parallel to the plane in which the teeth would have met, is what we would expect to see if



**Figure 1 | The evolution of mammalian middle ears.** Wang *et al.*<sup>1</sup> report the discovery of a fossil of a previously unknown mammalian species, *Jeholbaatar kielanae*. Its middle ear is similar to that of an extinct animal called *Arboroharamiya*. **a**, This similarity might indicate that *Jeholbaatar* and *Arboroharamiya* should be grouped close together on a mammalian family tree, and suggests that the 'palinal' chewing motion used by *Jeholbaatar* and *Arboroharamiya* has a single origin in a shared ancestor. Also shown in this tree are platypuses (*Ornithorhynchus*) and opossums (*Didelphis*), mammals that don't use palinal chewing and that have middle-ear configurations that are distinct from each other and from *Jeholbaatar* and *Arboroharamiya*. **b**, However, there is some debate about whether *Arboroharamiya* were mammals. If not, as in this tree, then the similar middle ears of *Jeholbaatar* and *Arboroharamiya* evolved independently. **c**, The configurations of the left middle-ear bones of these four creatures are presented as viewed directly from above, with the animal's front to the right. The different configurations of the incus, malleus and surangular bones might reflect the evolution of jaw specializations before bones separated from the jaw to form the ear. (Images based on ref. 1 and not shown to scale.)

palinal chewing had evolved before the middle ear was separate from the jaw<sup>4</sup>.

During transitional evolutionary stages, when the malleus was connected to the mandible, palinal jaw movement would have constrained the plane in which the malleus and incus could have been in contact; had the incus been in the more familiar posterior position found in most mammals today, it would have acted as a stop on backward jaw motion. Once palinal motion for chewing was established, increasing the distance the lower jaw moved forwards and backwards on the jaw joint would have made chewing more efficient. Any remaining tether to the ear would have limited the distance that the lower jaw could travel in a single chew, so selection pressure for a fully separate ear and jaw would have been strong, and full separation could have evolved rapidly.

The other animal known to have a surangular in the ear is *Arboroharamiya*, a member of an ancient group known as euharamiyidans with a palinal element to its chewing and an earlier origin than that of multituberculates<sup>4,5</sup>. *Arboroharamiya*, like *Jeholbaatar*, has its incus positioned above the malleus<sup>4,6</sup>. The relationship between euharamiyidans and multituberculates on the evolutionary tree is a matter of lively debate, with some studies, including that of Wang and colleagues, showing them to be closely related within mammals<sup>3,4,7</sup>, whereas others place euharamiyidans on a lineage that branched off before the common ancestor of living mammals evolved<sup>8,9</sup>. If the latter scenario is the case, then euharamiyidans would represent a fourth instance of the independent evolution of a fully detached middle ear.

The question of whether the similarities between the ears of *Jeholbaatar* and *Arboroharamiya* reflect a close relationship on the evolutionary tree or independent (convergent) evolution driven by similar chewing adaptations is further complicated by another consideration: the incus of living platypuses (*Ornithorhynchus*) and echidnas, or spiny anteaters (*Tachyglossus*), also lies above the malleus. These mammals belong to a group called monotremes, whose middle ear evolved independently of that of other mammals. Monotremes do not use a palinal chewing motion, and the teeth of fossil monotremes do not suggest that such a motion occurred in early members of that lineage<sup>10</sup>. They might have this arrangement of their incus and malleus for reasons that are entirely different from those explaining the arrangement of these bones in multituberculates or euharamiyidans. Monotremes do not retain a recognizable surangular. If the similarities in the middle ears of *Jeholbaatar* and *Arboroharamiya* reflect the functional similarity in the way the animals chewed, the unfused surangular in *Jeholbaatar* and *Arboroharamiya* might simply

reflect the rapidity with which the transition to detachment of the middle ear from the jaw occurred, spurred on by the increased efficiency in food processing that this complete separation would have provided.

**Anne Weil** is in the Department of Anatomy and Cell Biology, Oklahoma State University Center for Health Sciences, Tulsa, Oklahoma 74107, USA.  
e-mail: anne.weil@okstate.edu

1. Wang, H., Meng, J. & Wang, Y. *Nature* **576**, 102–105 (2019).

2. Weil, A. & Krause, D. W. in *Evolution of Tertiary Mammals of North America* Vol. 2 (eds Janis, M., Gunnell, G. F. & Uhen, M. D.) Ch. 2 (Cambridge Univ. Press, 2008).
3. Meng, J., Wang, Y. & Li, C. *Nature* **472**, 181–185 (2011).
4. Han, G., Mao, F., Bi, S., Wang, Y. & Meng, J. *Nature* **551**, 451–456 (2017).
5. Mao, F. & Meng, J. *Palaeontology* **62**, 639–660 (2019).
6. Meng, J. et al. *J. Anat.* <https://doi.org/10.1111/joa.13083> (2019).
7. Bi, S., Wang, Y., Guan, J., Sheng, X. & Meng, J. *Nature* **514**, 579–584 (2014).
8. Luo, Z.-X. et al. *Nature* **458**, 326–329 (2017).
9. Huttenlocker, A. K., Grossnickle, D. M., Kirkland, J. I., Schultz, J. A. & Luo, Z.-X. *Nature* **558**, 108–112 (2018).
10. Rich, T. H. et al. *Acta Paleontol. Polon.* **46**, 113–118 (2001).

This article was published online on 27 November 2019.

## Condensed-matter physics

# Electrons in graphene go with the flow

**Klaus Ensslin**

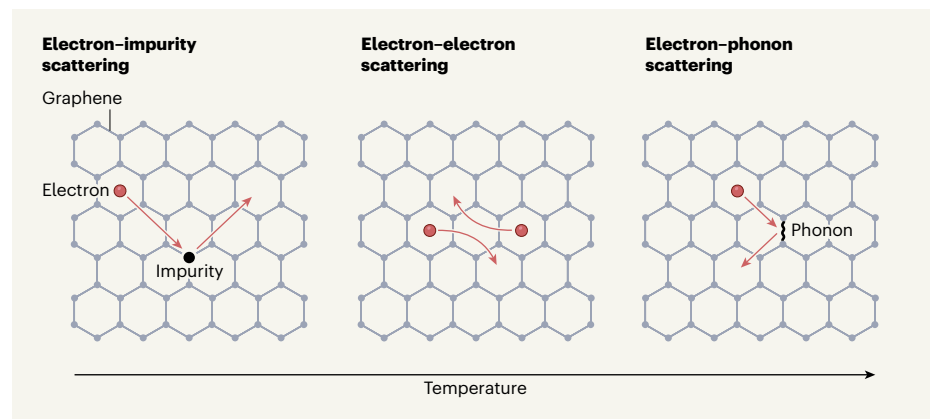
Scattering between electrons in the material graphene can cause these particles to flow like a viscous liquid. Such flow, which has previously been detected using measurements of electrical resistance, has now been visualized. **See p.75**

Water in a river shows a variety of flow patterns and whirls. Any obstacle in the river, such as a bridge pillar or simply a rough bank, will lead to a distinctive flow pattern. It has been comparatively less obvious how electrons flow in a solid. But on page 75, Sulpizio *et al.*<sup>1</sup> report an experiment in which the flow pattern of electrons in an electrical conductor is imaged.

The electrical resistance of a metal is caused by electrons being scattered from impurities in the material's atomic lattice or from lattice vibrations called phonons. However, it is not

affected by electron–electron scattering. When two electrons scatter off each other, their individual momenta are changed by the scattering event. But the total momentum of the two electrons is conserved, as is the total momentum of a sea of electrons in a metal. Therefore, simply measuring the resistance of a metal will not unveil the effects of electron–electron scattering.

To nail down these effects, materials need to be tuned to a regime in which electron–electron scattering is dominant and the



**Figure 1 | Electron interactions in graphene.** The material graphene consists of a single layer of carbon atoms arranged in a hexagonal lattice. Electrons flowing through graphene can be scattered from impurities (such as foreign atoms in the lattice), from other electrons and from lattice vibrations known as phonons. At low temperatures, electron–impurity scattering dominates. By contrast, at high temperatures, electron–phonon scattering takes over. Sulpizio *et al.*<sup>1</sup> report observations of graphene at intermediate temperatures for which the rate of electron–electron scattering is the largest among all scattering rates.



electrons flow like a viscous liquid<sup>2,3</sup>. At low temperatures, electron–electron (as well as electron–phonon) scattering is suppressed and electron–impurity scattering dominates. Conversely, at high temperatures, electron–phonon scattering takes over. For graphene (a single layer of carbon atoms arranged in a honeycomb lattice), there is an intermediate temperature range<sup>4</sup> (50–250 kelvin) for which the rate of electron–electron scattering is the highest among all scattering rates (Fig. 1). However, even in this case, the material’s resistance will not be modified by electron–electron scattering because of momentum conservation.

One way to investigate the viscous-flow regime has been to measure a local resistance, known as vicinity resistance<sup>4</sup>, on an extremely small scale. The value of this quantity changes sign in the case of viscous flow. Another option has been to observe an effect called superballistic resistance<sup>5</sup> for electrons flowing through a narrow opening in a material. Here, the resistance is reduced below the value expected for a ballistic system, in which there is effectively no scattering. Such pioneering experiments were crucial for demonstrating that viscous electron flow can be important in electron transport. However, they provide only indirect evidence for the existence of such flow and do not give insights into the spatial arrangements of flow patterns.

Electrons passing through a sample of a conducting material are driven by an electric field. As a result, there is a voltage gradient along the direction of current flow. Unfortunately, this local voltage gradient is independent of the flow regime. But when a weak magnetic field is applied to the sample, another voltage, known as a Hall voltage, is produced perpendicular to the direction of current flow. The spatial profile of the Hall voltage does provide information about the flow characteristics.

Sulpizio and colleagues use a sensitive electric-field sensor that enables local probing of this Hall voltage. The sensor is an innovative technology developed by this research group<sup>6</sup>. It consists of an electronic device called a single-electron transistor, the conductance of which depends sensitively on its electrostatic environment.

In the present work, the sensor is made from ultraclean carbon nanotubes. Individual electrons are confined within these nanotubes by electrodes. Such an arrangement provides the required sensitivity for detecting weak electric fields or voltage gradients, such as those associated with the Hall voltage. The spatial resolution of the sensor is limited by its size and the distance of the sensor to the object to be probed.

Changing the temperature and the number of charge carriers per given area in the sample induces different flow regimes, which lead to

different Hall-voltage profiles. Sulpizio *et al.* use this property to image local electric fields in a uniform layer of graphene, and investigate the transition between the regime in which electron–electron scattering dominates and those in which electron–phonon or electron–impurity scattering takes over.

The authors demonstrate experimentally how electron–electron scattering alters the Hall-voltage profile of a uniform conductor. Viscous flow in liquids leads to turbulence and whirls, depending on the viscosity of the liquid and on obstacles to the flow. However, the observation of such features in electron transport is beyond the scope of the present work and could require different experimental tools, such as sensitive magnetic-field sensors, or samples that have complex geometries.

What do Sulpizio and colleagues’ results mean for our understanding of electron transport in conductors? In the viscous regime, the flow of electrons is described by a universal hydrodynamic concept known as Poiseuille flow. The authors’ imaging of electronic Poiseuille flow is a breakthrough in the study of electron transport as well as a demonstration of a sophisticated imaging technique that combines high spatial resolution with extreme sensitivity. We now know that electron flow can be diffusive, ballistic or viscous, and that there are experimental tools for differentiating between these regimes.

For solid-state systems in general, electron–electron interactions are relevant for phenomena as diverse as ferromagnetism (the familiar type of magnetism found in iron bar magnets) and the fractional quantum Hall effect (whereby electrons in a strong magnetic field act together to behave like particles that have a fractional electric charge). The authors’ technique could also be used to investigate, on a local scale, the superconductivity that was discovered last year in a twisted bilayer of graphene<sup>7</sup>. The potential to extract local information about strongly interacting systems of electrons will have far-reaching consequences for this field. Further applications of the technique could enable local probing of electric fields as they arise in complex quantum circuits – which might one day lead to a quantum computer.

**Klaus Ensslin** is in the Laboratory for Solid State Physics, ETH Zurich, 8093 Zurich, Switzerland.  
e-mail: ensslin@phys.ethz.ch

1. Sulpizio, J. A. *et al.* *Nature* **576**, 75–79 (2019).
2. Andreev, A. V., Kivelson, S. A. & Spivak, B. *Phys. Rev. Lett.* **106**, 256804 (2011).
3. Levitov, L. & Falkovich, G. *Nature Phys.* **12**, 672–676 (2016).
4. Bandurin, D. A. *et al.* *Science* **351**, 1055–1058 (2016).
5. Kumar, R. K. *et al.* *Nature Phys.* **13**, 1182–1185 (2017).
6. Ella, L. *et al.* *Nature Nanotechnol.* **14**, 480–487 (2019).
7. Cao, Y. *et al.* *Nature* **556**, 43–50 (2018).

## Neurodevelopment

# Birth of a motor circuit visualized

**Kristen P. D’Elia & David Schoppik**

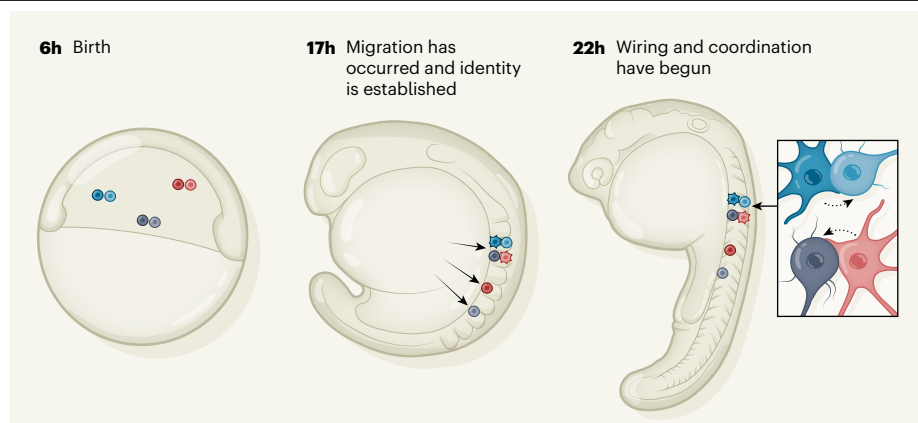
A sophisticated imaging pipeline has been developed to track neurons in early-stage zebrafish embryos over time and space. It reveals how newborn neurons come together to build a spinal cord capable of locomotion.

Where a person comes from and what they do are often considered key parts of their identity. Similarly, neurons can be categorized by both their developmental history and their role in the nervous system. But, just as knowing someone’s job title does not necessarily tell you what part they play in a team at work, knowing what role a neuron has does not mean that we understand how it comes together with other diverse neuron types to form circuits – for instance, to permit movement. Writing in *Cell*, Wan *et al.*<sup>1</sup> describe an imaging protocol that will help researchers determine how neural circuits form. They use their method to comprehensively chart

motor-circuit assembly and emerging function in the spinal cord of zebrafish.

In vertebrate embryos, the first neuronal circuits to respond to sensory information and orchestrate movement are found in the spine<sup>2</sup>. These motor circuits are assembled from dozens of molecularly specialized types of neuron. Nonetheless, this is a relatively simple set-up, making it a useful system for studying how neuronal circuits come together to produce behaviour – in this case, muscles contracting in distinct patterns.

Wan *et al.* set out to study the formation of these early motor circuits in zebrafish embryos (Fig. 1). This research group has long



**Figure 1 | Tracking the building blocks of a circuit.** Wan *et al.*<sup>1</sup> have developed an imaging and computational pipeline to track neurons of the zebrafish spinal cord, from their ‘birth’ 6 hours after embryo fertilization until they begin to show the coordinated activity of a motor circuit at 22 hours. The authors traced newly born sister cells (derived from the same immediate ancestor, indicated by different shades of the same colour). By 17 hours, the cells have migrated to their mature positions and adopted molecular characteristics of either motor neurons (star-shaped cell) or interneurons (circular cell). By 22 hours, the cells become wired into coordinated circuits (inset). Motor neurons are the first to become active, and the authors showed that they then imprint their activity onto other neurons (dotted arrows), leading these neurons to adopt the same activity pattern.

been at the forefront of *in vivo* microscopy, pioneering light-sheet microscopy techniques that can illuminate all of the individual cells that make up developing organisms such as zebrafish without harming them. Zebrafish are well suited to such studies because they are small, transparent and develop rapidly.

The researchers imaged zebrafish from 6 hours after embryo fertilization, when spinal neurons first arise from their progenitors, to 22 hours after fertilization, when the patterns of neuronal activity that trigger tail movements begin. The imaging process generated vast libraries of images that Wan and colleagues processed to extract information about the location of individual cells over time. In addition, the authors optimized their microscope design to allow them to measure emergent patterns of functional activity from individual cells. The result was a data set that enabled the group to track the organization and function of every cell in the zebrafish spinal cord throughout early development.

Motor neurons and interneurons are key neuron types in spinal motor circuits. The former are responsible for triggering muscle-fibre contraction and the latter coordinate signalling within and between circuits<sup>3</sup> (for example, to ensure alternating left–right movements during swimming). Motor neurons have often been thought of as passive cells controlled by upstream interneuron inputs, whereas interneurons had been thought to be the driving force behind the assembly and function of spinal motor circuits<sup>4</sup>. But over the past few years, evidence has emerged that both developing<sup>5</sup> and mature<sup>6</sup> motor neurons can control their connections to interneurons, and

even control interneuron activity. In zebrafish, motor neurons are the first spinal neurons to display spontaneous activity patterns<sup>7</sup>. As a circuit develops, neurons often first become active on their own, and then coordinate their activity with that of other neurons. Wan *et al.* therefore asked whether this activity originates in the motor neurons themselves, or reflects interneuron control.

The authors found that select motor neurons seem to impose their own activity on neighbouring motor neurons and interneurons, producing pairs of cells that have the same activity patterns. Thus, the earliest patterns of collective activity are initiated by motor neurons. This finding adds to the emerging picture of motor neurons as a fundamental driver of spinal-cord development. Consistent with previous findings, the authors also confirmed that interneurons coordinate the global patterns of activity necessary at later developmental stages for tail movement.

One theory of neural development states that cells that have a shared ancestry are destined to have common connectivity, and to perform similar roles in a circuit<sup>8</sup>. Evidence for such determinism remains contentious, reflecting the challenge of tracing related neurons as they migrate<sup>9</sup>. But Wan and colleagues were able to investigate this issue, thanks to their ability to comprehensively track cells over time.

The authors examined the activity of sister neurons — those that shared an immediate ancestor. In line with ideas of determinism, sister neurons that ended up in close proximity to one another were more likely than unrelated neurons to be co-active. But, intriguingly, most sibling pairs did not remain

close to one another. Indeed, sister neurons were just as likely to migrate to opposite sides of the spinal cord, where they would participate in different phases of movement. Thus, ancestry can explain only a small part of functional organization. That said, Wan and colleagues’ study is limited to the earliest part of development, well before zebrafish hatch and swim freely. It will be interesting to re-evaluate questions of ancestral determinism over longer periods of time.

Another limitation of the authors’ technique is that their cutting-edge microscope is best suited to small model organisms. It would be interesting to analyse whether their findings also apply to more-complex organisms. However, current microscopes cannot be used for such purposes.

Notably, the group that performed the study (and the Janelia Research Campus in Ashburn, Virginia, at which it works) is committed to providing access to the microscope used in the current work. In addition, the authors’ data and analysis pipelines are available to download. Thus, other researchers can further assess the relationship between the developmental history and function outlined in the current study.

Advances in the transcriptional profiling of single cells have revealed remarkable variability among neurons<sup>10</sup>, making circuit development ever-more fascinating but incredibly challenging to fully understand. Until we have a greater understanding of the molecular logic that enables neurons to form motor circuits, our ability to prevent, diagnose and treat disorders of movement will remain limited. The apparatus and analysis pipeline developed by Wan *et al.* present a technically demanding but demonstrably fruitful path towards better grasping how a neuron’s birth shapes its future role in a circuit.

**Kristen P. D’Elia** and **David Schoppik** are in the Departments of Otolaryngology and of Neuroscience and Physiology, Neuroscience Institute, New York University Langone Medical Center, New York, New York 10016, USA. e-mail: schoppik@gmail.com

1. Wan, Y. *et al.* *Cell* **179**, 355–372 (2019).
2. Hanson, M. G. & Landmesser, L. T. *J. Neurosci.* **23**, 587–600 (2003).
3. Gosgnach, S. *et al.* *J. Neurosci.* **37**, 10835–10841 (2018).
4. D’Elia, K. P. & Dasen, J. S. *Neural Dev.* **13**, 10 (2018).
5. Baek, M., Pivetta, C., Liu, J.-P., Arber, S. & Dasen, J. S. *Cell Rep.* **21**, 867–877 (2017).
6. Song, J., Ampatzis, K., Björnfors, E. R. & El Manira, A. *Nature* **529**, 399–402 (2016).
7. Warp, E. *et al.* *Curr. Biol.* **22**, 93–102 (2012).
8. Gao, P., Sultan, K. T., Zhang, X.-J. & Shi, S. H. *Development* **140**, 2645–2655 (2013).
9. Mayer, C. *et al.* *Neuron* **87**, 989–998 (2015).
10. Holguera, I. & Desplan, C. *Science* **362**, 176–180 (2018).

This article was published online on 20 November 2019.



## Genetic engineering

# CRISPR tool enables precise genome editing

Randall J. Platt

The ultimate goal of genome editing is to be able to make any specific change to the blueprint of life. A 'search-and-replace' method for genome editing takes us a giant leap closer to this ambitious goal. **See p.149**

Variation in the DNA sequences that constitute the blueprint of life is essential to the fitness of any species, yet thousands of DNA alterations are thought to cause disease. After decades of research in genetics and molecular biology, tremendous progress has been made in developing genome-editing tools for correcting such alterations. But a seemingly fundamental limit to the efficiency and precision of gene editing was reached, owing to the tools' reliance on complex and competing cellular processes. On page 149, Anzalone *et al.*<sup>1</sup> describe 'search-and-replace' genome editing, in which the marriage of two molecular machines enables the genome to be altered precisely. The technique has immediate and profound implications for the biomedical sciences.

Human efforts to engineer genomes pre-date knowledge of genes or even of the source of heredity. The first genome engineering relied on natural variation and artificial selection through selective breeding. Modern maize (corn), for example, was 'engineered' from its wild ancestor, teosinte, through artificial selection more than 9,000 years ago<sup>2</sup>. Later progress was fuelled by the realization that DNA sequences shape life, and that evolution can be augmented and artificially accelerated through the use of mutagenic agents, such as radiation or chemicals.

Next came the discovery that cellular processes for repairing mistakes in DNA sequences could be hijacked, allowing sequences from a foreign 'template' DNA to be inserted into the genome at DNA breaks<sup>3</sup>. This process is greatly enhanced if the DNA is intentionally damaged<sup>4,5</sup> – a finding that sparked a search of more than 20 years for an enzyme that could specifically cut DNA at locations of interest. The search culminated in the adoption of the bacterial CRISPR–Cas9 system, in which the enzyme Cas9 uses a customizable RNA guide to search for DNA sequences to cut in human cells<sup>6–8</sup> (Fig. 1a).

CRISPR–Cas9 sparked a revolution in the biomedical sciences by making genome editing accessible to all researchers, but,

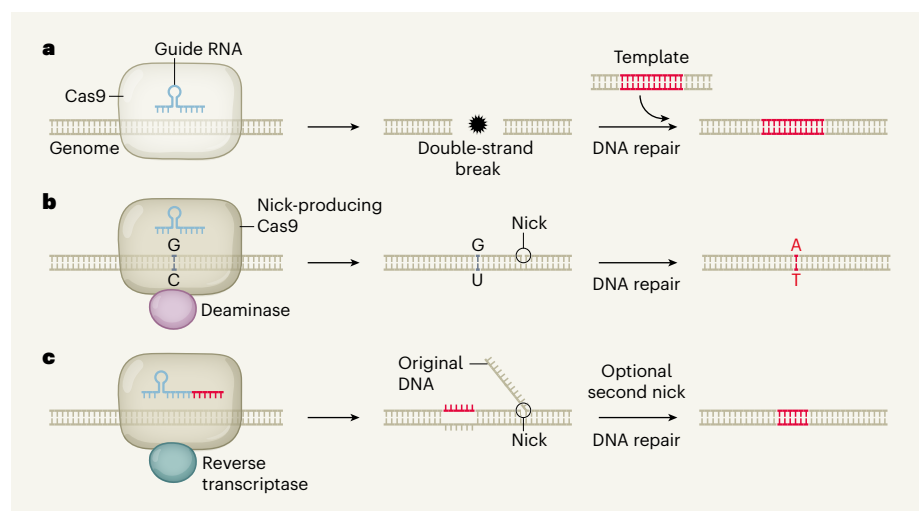
ultimately, it is just a fancy pair of molecular scissors that cuts DNA. Because cuts in DNA are deadly to cells, they are urgently repaired by one of many independent pathways. In the context of genome editing, the desired outcome is for repair to be directed by a template DNA, resulting in precise edits. But most cells prefer to use an alternative mechanism, in which the DNA template is ignored and the two broken ends of DNA are imperfectly stitched back together – a major limitation for genome editing.

Much effort over the past few years has focused on shifting the balance from imperfect to precise editing. One effective strategy is to edit DNA without cutting both DNA strands

in the helix – double-strand breaks are the main insult that leads to imperfect edits. A milestone in this regard was the development of base editing<sup>9</sup>, a process in which a version of the Cas9 enzyme that cuts only one DNA strand is combined with an enzyme that can switch one specific DNA base for another, near the nick site (Fig. 1b). However, the technical constraints of base editing, and the need to modify more than just single DNA bases, meant that new genome-editing approaches were still desperately needed.

Anzalone and co-workers now largely fill this need with a technique called prime editing. Their approach relies on a hybrid molecular machine consisting of a modified version of Cas9, which cuts only one of the two DNA strands, and a reverse transcriptase enzyme, which installs new and customizable DNA at the cut site (Fig. 1c). This marriage parallels a naturally occurring process in yeast, in which DNA that corresponds to an RNA sequence is incorporated into the genome by a reverse transcriptase<sup>10</sup>.

The prime-editing process is orchestrated by an engineered, two-part RNA guide. The 'search' part of the guide directs Cas9 to a specific sequence in the DNA target, where it cuts one of the two DNA strands. The reverse transcriptase then produces DNA complementary to the sequence in the 'replace' part of the RNA guide, and installs it at one of the cut DNA ends, where it takes the place of the original DNA sequence.



**Figure 1 | Evolution of genome editing.** **a**, In conventional genome editing, a Cas9 enzyme is directed to a position in the genome by a guide RNA, and produces a double-strand break. The host cell's DNA-repair machinery fixes the break, guided by a template DNA, incorporating template sequences into the duplex. **b**, In an approach called base editing, a Cas9 that produces only single-strand breaks (nicks) works with a deaminase enzyme. The deaminase chemically modifies a specific DNA base – here, a cytidine base (C) is converted to uracil (U). DNA repair then fixes the nick and converts a guanine–uracil (G–U) intermediate to an adenine–thymine (A–T) base pair. This method is more precise than **a**, but makes only single-nucleotide edits. **c**, Anzalone *et al.*<sup>1</sup> report prime editing, which can precisely edit DNA sequences. A nick-producing Cas9 and a reverse transcriptase enzyme produce nicked DNA into which sequences corresponding to the guide RNA have been incorporated. The original DNA sequence is cut off, and DNA repair then fixes the nicked strand to produce a fully edited duplex. In some cases, another nick is made in the unedited strand of the duplex before the DNA-repair step (not shown).

At this point, the duplex DNA being modified consists of two non-complementary strands: the edited strand, and the intact strand that wasn't cut by Cas9. Non-complementary sequences are not tolerated in cells, so one of the strands must be fixed by DNA-repair processes to match the other, with the intact strand typically being preferentially retained. The authors therefore usually had to use a second RNA guide to direct a cut to the intact strand, to increase the chances that that strand would be repaired to match the edited sequence. The cut must be made strategically to avoid breaking both strands at the same time or place.

Anzalone *et al.* demonstrate the versatility of prime editing by using it to efficiently and precisely install a wide range of sequences into DNA. For example, they used it *in vitro* in human embryonic kidney cells to correct the mutations that give rise to the blood disorder sickle-cell disease, and to edit the mutations that cause the neurological condition Tay–Sachs disease. Imperfect edits were almost entirely avoided. The authors also carried out edits in human cancer cells and in mouse neurons *in vitro*.

For decades, the potential of genome editing has been constrained by the difficulty of making precise modifications, and so applications have focused heavily on situations in

which imperfect DNA edits are useful. For example, such edits can be used to impair the function of a gene, providing an avenue for understanding its function<sup>11</sup>. Prime editing now makes it faster and easier than before to install or correct one or many specific mutations (such as those found in human patients, or synthetic sequences that are useful for research purposes). And it makes more cell types available for manipulation than was previously possible. The chains that have shackled gene editing have thus come off – no doubt quickening the pace of research and enabling a list of new applications.

Nevertheless, prime editing has limitations. First, the sophisticated, multi-step molecular dance that occurs between the prime-editing components is not yet predictable and doesn't always turn out as intended. Imperfect random edits can therefore still arise, which means that several combinations of components might need to be tested, to work out the choreographies required for each edit of interest. Second, delivering the large prime-editing system into some cell types could be challenging, given that many previous attempts have faltered with the conventional Cas9 system<sup>12</sup>, which is roughly half the size.

For research purposes, these limitations are mostly just inconvenient, and will probably be overcome through follow-up work directed

at better understanding and fine-tuning the method. For medical applications, however, these issues present a much greater challenge – imperfect DNA edits are unacceptable, and efficient delivery of the prime-editing system to cells will be crucial. So although prime editing certainly has the potential to give us unprecedented control over the blueprint of life, only time will tell whether it becomes just another tool in the CRISPR toolbox or a cure-all for genetic diseases.

**Randall J. Platt** is in the Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland.  
e-mail: rplatt@ethz.ch

1. Anzalone, A. V. *et al.* *Nature* **576**, 149–157 (2019).
2. Matsuoka, Y. *et al.* *Proc. Natl Acad. Sci. USA* **99**, 6080–6084 (2002).
3. Capecchi, M. R. *Science* **244**, 1288–1292 (1989).
4. Rudin, N., Sugarmann, E. & Haber, J. E. *Genetics* **122**, 519–534 (1989).
5. Rouet, P., Smih, F. & Jasin, M. *Mol. Cell. Biol.* **14**, 8096–8106 (1994).
6. Jinek, M. *et al.* *Science* **337**, 816–821 (2012).
7. Cong, L. *et al.* *Science* **339**, 819–823 (2013).
8. Mali, P. *et al.* *Science* **339**, 823–826 (2013).
9. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. *Nature* **533**, 420–424 (2016).
10. Keskin, H. *et al.* *Nature* **515**, 436–439 (2014).
11. Platt, R. J. *et al.* *Cell Rep.* **19**, 335–350 (2017).
12. Cox, D. B. T., Platt, R. J. & Zhang, F. *Nature Med.* **21**, 121–131 (2015).

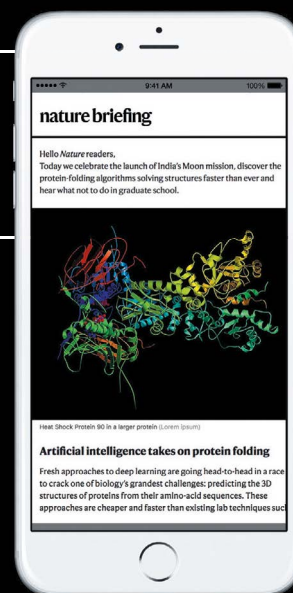
This article was published online on 11 November 2019.

# nature briefing

**What matters in science and why – free in your inbox every weekday.**

The best from *Nature's* journalists and other publications worldwide. Always balanced, never oversimplified, and crafted with the scientific community in mind.

**SIGN UP NOW**  
[go.nature.com/briefing](https://go.nature.com/briefing)



**nature**



# The integrative biology of type 2 diabetes

<https://doi.org/10.1038/s41586-019-1797-8>

Received: 27 March 2019

Accepted: 18 September 2019

Published online: 4 December 2019

Michael Roden<sup>1,2,3\*</sup> & Gerald I. Shulman<sup>4\*</sup>

Obesity and type 2 diabetes are the most frequent metabolic disorders, but their causes remain largely unclear. Insulin resistance, the common underlying abnormality, results from imbalance between energy intake and expenditure favouring nutrient-storage pathways, which evolved to maximize energy utilization and preserve adequate substrate supply to the brain. Initially, dysfunction of white adipose tissue and circulating metabolites modulate tissue communication and insulin signalling. However, when the energy imbalance is chronic, mechanisms such as inflammatory pathways accelerate these abnormalities. Here we summarize recent studies providing insights into insulin resistance and increased hepatic gluconeogenesis associated with obesity and type 2 diabetes, focusing on data from humans and relevant animal models.



**Anniversary  
collection:**  
[go.nature.com/  
nature150](http://go.nature.com/nature150)

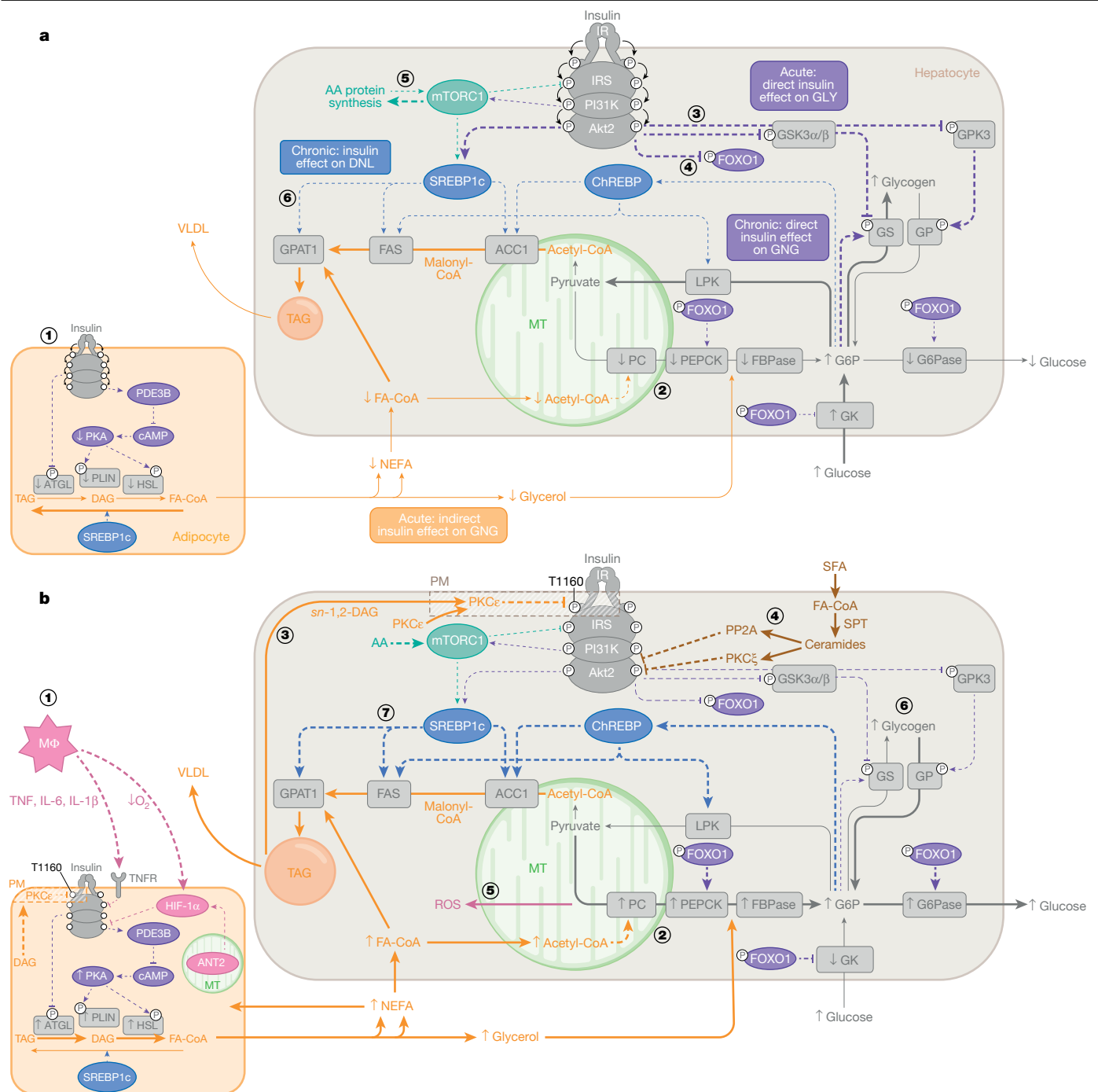
Over the past 50 years, the prevalence of diabetes mellitus has continued to increase, spreading from western countries to the western Pacific, Asia and Africa. Current projections estimate an increase of more than 50% between 2017 and 2045, leading to around 693 million people suffering from diabetes, with estimated healthcare costs of about US\$850 billion per year<sup>1</sup>. This epidemic mainly results from an increase in the incidence of type 2 diabetes (T2D), a heterogeneous disease characterized by deficient insulin secretion by pancreatic islet  $\beta$ -cells in the context of impaired insulin sensitivity, termed insulin resistance. A genome-wide association study (GWAS) found more than 400 T2D-associated gene variants—mostly related to islet function, but the roles of the individual genes are minor and explain less than 20% of overall disease risk<sup>2</sup>. Lifestyle-modification studies demonstrating T2D remission underline the predominant role of acquired alterations<sup>3,4</sup>, including intake of highly palatable, energy-dense refined food, sedentary behaviour and other factors (for example, environmental pollution, socioeconomic and psychosocial conditions, smoking and sleep deprivation)<sup>5</sup>. Moreover, parental lifestyle, intrauterine programming and early postnatal metabolic alterations may influence the risk profile<sup>6</sup> via DNA methylation<sup>7</sup>. The roles of these mechanisms and of the gut metagenome are controversial in humans and beyond the scope of this Review. This Review focuses on studies in humans and relevant rodent models to provide an outlook on future precision medicine for T2D by better understanding its pathogenesis.

## Fed-to-fasting transition and insulin resistance

Fasted humans display impaired insulin-stimulated glucose disposal and elevation of certain, mainly branched-chain, amino acids

and nonesterified fatty acids (NEFA) in plasma despite low-to-normal glycaemia and hypoinsulinaemia<sup>8</sup>. While initially hepatic glycogenolysis and gluconeogenesis maintain normoglycaemia<sup>9</sup>, the shift from carbohydrate to fatty acid oxidation preserves glucose for obligate glucose utilizers (such as the brain, red blood cells and renal medulla) and essential protein stores, which would otherwise be used for gluconeogenesis (Fig. 1a). Stimulation of gluconeogenesis has mostly been attributed to decreasing plasma insulin and increasing plasma glucagon concentrations favouring gluconeogenic enzyme transcription. Recent studies in rats have demonstrated a critical role of the leptin–hypothalamic–pituitary–adrenal (HPA) axis in mediating the fed-to-fasting transition via glucocorticoid regulation of white adipose tissue (WAT) lipolysis<sup>10–13</sup>—similar to mechanisms that operate in uncontrolled diabetes<sup>14</sup>. These studies show that the early postabsorptive decline in hepatic glycogenolysis is predominantly responsible for the fall in plasma insulin and glucose concentrations, resulting in approximately 50% reduction in plasma leptin concentrations. Thus, leptin acts as important fuel gauge for energy stored as triacylglycerol (TAG) in WAT and as glycogen in liver, signalling to the brain when both energy depots are depleted<sup>10</sup>. The fall in leptin to less than 1 ng ml<sup>-1</sup> stimulates the HPA axis, ultimately increasing plasma corticosterone concentrations, which, during hypoinsulinaemia, stimulates WAT lipolysis, and release of NEFA and glycerol, with a switch towards fatty acid oxidation. Increased fatty acid flux to the liver increases hepatic  $\beta$ -oxidation and acetyl-CoA content (Fig. 1b). This allosterically activates pyruvate carboxylase flux, which, along with increased glycerol flux to the liver, is essential for maintaining hepatic gluconeogenesis and endogenous glucose production (EGP) during starvation. Simulation of fasting conditions also increased the contribution of gluconeogenesis to EGP by up to 75%, probably owing to lipid-dependent control of hepatic glycogen stores<sup>15</sup>. Starvation also promotes hepatic accumulation of TAG and diacylglycerol (DAG)<sup>10</sup>, which can occur independently of the direct action of hepatic insulin on de novo lipogenesis (DNL)<sup>16</sup>. Subsequently, the novel protein kinase C isoform  $\epsilon$  (PKC $\epsilon$ ) is translocated to the plasma membrane, where it binds to and phosphorylates Thr1160 of the insulin receptor (IR), thereby inhibiting IR kinase activity<sup>17</sup> (Fig. 1b). Prolonged

<sup>1</sup>Division of Endocrinology and Diabetology, Medical Faculty, Heinrich-Heine University, Düsseldorf, Germany. <sup>2</sup>Institute for Clinical Diabetology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich-Heine University, Düsseldorf, Germany. <sup>3</sup>German Center for Diabetes Research, Partner Düsseldorf, Düsseldorf, Germany. <sup>4</sup>Departments of Internal Medicine and Cellular and Molecular Physiology, Yale Diabetes Research Center, Yale School of Medicine, New Haven, CT, USA. \*e-mail: michael.roden@ddz.de; gerald.shulman@yale.edu



fasting also resulted in 60% lower rates of glucose–alanine cycling, with a 50% reduction in hepatic mitochondrial oxidation, demonstrating an interorgan link between liver and muscle during the fed-to-fasting transition in both rats<sup>10</sup> and humans<sup>18,19</sup>. Unbiased metabolomic analysis suggests that the same sequence of events occurs in ten-day fasted humans and reveals discrete starvation phases with gluconeogenic amino acid consumption and subsequent surge in lipids with a high degree of unsaturation and chain length<sup>12,13</sup>, reflecting increased adipocyte NEFA release. This study also reported a rapid fall of around 50% in circulating leptin and an early rise in plasma glucocorticoids, similar to that occurring in fasting<sup>20</sup> and anorexia nervosa<sup>21</sup>. Although circulating leptin tightly reflects WAT mass, fasting-induced hypoleptinaemia can occur independently and initiate neuroendocrine adaptation<sup>22,23</sup>.

This mechanism may have been important for survival during periods of famine and explain the high evolutionary conservation of the

Thr1160 residue in the catalytic loop of the IR. Likewise, the blind cave fish *Astyanax mexicanus*—which endures infrequent limited nutrient availability—develop hyperglycaemia, steatosis and insulin resistance owing to a mutation in its IR gene<sup>24</sup>, similar to the one observed in some patients with Rabson–Mendenhall syndrome<sup>25</sup>. Despite abnormal energy metabolism, these fish show delayed senescence, further supporting the survival benefit of limited insulin-dependent glucose disposal. Thus, DAG and novel PKC (nPKC)-induced insulin resistance may have served a key evolutionary role to promote survival during starvation, while favouring metabolic syndrome and T2D during overnutrition.

### Transition from normoglycaemia to hyperglycaemia

Longitudinal studies have demonstrated that people who later develop T2D display gradually increasing fasting and postprandial



**Fig. 1 | Adipose–liver interaction under insulin-sensitive and insulin-resistant conditions.** **a**, Under physiological postprandial conditions, insulin rapidly stimulates lipid storage by inhibiting lipolysis via adipose triglyceride lipase (ATGL and CGI-58), phosphodiesterase 3B (PDE3B) and protein kinase A (PKA)-controlled hormone-sensitive lipase (HSL) and perilipins (PLINs), and stimulating lipogenesis (1). Lower NEFA–glycerol flux decreases hepatic acetyl-CoA and glycerol, thereby acutely diminishing gluconeogenesis (GNG), reflecting indirect insulin action (2). Direct hepatic IR activation stimulates glycogen synthesis (GLY) (3) and chronic transcriptional control via decreased FOXO1 to downregulate gluconeogenic enzymes and upregulate glucokinase (GK), increasing glucose-6-phosphate (G6P) levels (4). Inhibition of serine phosphorylation of GSK3 activates and increases glycogen synthase (GS) flux (3). Furthermore, glucose-6-phosphate allosterically activates glycogen synthase and inhibits glycogen phosphorylase (GP), resulting in glycogen storage and suppressed glucose production. Insulin further participates in protein synthesis via mTORC1 (5) and DNL via FOXO1, carbohydrate and sterol response element binding proteins (ChREBP and SREBP1c) (6). **b**, In insulin-resistant states (obesity and T2D), adipocyte dysfunction—for example, owing to relative hypoxia induced by saturated fat-stimulated mitochondrial ANT2 triggering the transcription factor HIF-1 $\alpha$ —leads to chemokine secretion, attracting macrophages (M $\Phi$ ) (1). Immune cell infiltration inhibits adipocyte insulin sensitivity by mechanisms that ultimately increase lipolysis and NEFA

and glycerol flux to the liver. Here, acetyl-CoA allosterically stimulates pyruvate carboxylase flux (PC), subsequently raising fasting glucose production (2). Increased lipid re-esterification generates DAG, thereby activating PKC $\epsilon$  translocation and inhibitory Thr1160 phosphorylation of the IR (3) and increasing production of ceramides (4) and ROS (5), collectively promoting insulin resistance. Inhibiting direct hepatocellular insulin action acutely favours glycogenolysis and chronically upregulates gluconeogenesis with postprandial and finally continuous hyperglycaemia (6). In parallel, hepatic TAG deposition increases, not only from augmented lipid availability and DNL, partly controlled by insulin and FOXO1, but also via nutrient-sensitive pathways (ChREBP, SREBP1c and mTORC1) (7). Finally, endoplasmic reticulum (ER)-derived factors (PKR-like eukaryotic initiation factor 2 $\alpha$  kinase (PERK), inositol-requiring enzyme 1 (IRE1) and activating-transcription factor 6 (ATF6)) can induce an unfolded protein response, which may stimulate lipogenesis, supported by X-box binding protein 1 (XBP1) and inflammatory pathways. All of these mechanisms accelerate TAG accumulation and NAFLD progression. Dotted lines represent regulation, that is, stimulation or inhibition; thicker lines represent pathways with increased flux, thinner lines represent pathways with decreased flux. FAS, fatty acid synthase; PEPCK, phosphoenolpyruvate carboxykinase; FBPAse, fructose-1,2-bis-phosphatase; LPK, L-pyruvate kinase. P, phosphorylation.

glycaemia<sup>26–28</sup>. Insulin sensitivity, which is predominately dependent on age, sex and weight gain, declines decades before T2D onset, represents one of the earliest pathogenic events and can be mostly ascribed to reduced nonoxidative glucose metabolism<sup>29</sup> resulting from impaired insulin-stimulated storage of ingested carbohydrate as muscle glycogen<sup>30</sup> (Fig. 2b). Initially,  $\beta$ -cells compensate for insulin resistance by secreting more insulin, resulting in hyperinsulinaemia, which promotes hepatic DNL, steatosis, hyperlipidaemia and WAT expansion<sup>30</sup>. WAT dysfunction, due to insulin resistance and inflammation<sup>14,31,32</sup>, stimulates lipolysis, further aggravating hepatic insulin resistance and non-alcoholic fatty liver disease (NAFLD) (Fig. 1b). Additionally, increased NEFA and/or glycerol flux to the liver stimulates gluconeogenesis. Combined with declining  $\beta$ -cell function—at least partly due to glucolipotoxicity—typically occurring just before T2D onset<sup>27,28</sup>, this leads to fasting and postprandial hyperglycaemia. There appear to be population-specific differences (for example, quicker decline in  $\beta$ -cell function in African Americans<sup>33</sup> and increased hepatic lipid accumulation and muscle insulin resistance despite lower bodyweight in Asian Indians<sup>28,34</sup>). Nevertheless, without weight loss, insulin resistance and  $\beta$ -cell dysfunction occur simultaneously and continuously, increasing the risk of comorbidities even before glycaemia exceeds current criteria defining diabetes.

## Identifying distinct diabetes phenotypes

Recent studies challenge the traditional concept of T2D as single entity, as patients already exhibit a broad variability in insulin secretion and sensitivity at diagnosis<sup>35</sup>. Unbiased cluster analyses discriminated subgroups with different degrees of insulin deficiency and moderate obesity-related, moderate age-related or severe insulin resistance<sup>35,36</sup>. Whereas no known diabetes gene variants were associated with all clusters, a *TCF7L2* variant related to insulin deficiency and a *TM6SF2* variant related to the severely insulin-resistant cluster predicted nephropathy, cardiovascular disease<sup>36</sup> and NAFLD<sup>35</sup>. Soft clustering analyses point to further gene–phenotype associations<sup>37</sup> underlining different pathogenic mechanisms.

## Postprandial hepatic metabolite fluxes

Fasting hyperglycaemia in T2D results from increased rates of hepatic gluconeogenesis and EGP and from hepatic insulin resistance, characterized by reduced ability of insulin to suppress this process<sup>38–41</sup>.

This may be because of direct IR-mediated cell-autonomous or indirect effects (substrate availability, allosteric regulation or redox status)<sup>42</sup> (Fig. 1b). Recent studies showed that these indirect effects probably result from insulin action on WAT and mainly account for acute suppression of gluconeogenesis and EGP during postprandial hyperinsulinaemia<sup>14</sup>. Consistent with a minor role for direct hepatic effects of insulin, rodent models with altered hepatic insulin signalling exhibit relatively normal glucose tolerance and compensatory hyperinsulinaemia, with reduced hepatic glycogen synthesis as the only indication of disrupted insulin signalling<sup>14,43–47</sup>.

Direct assessment of glycogen synthesis by <sup>13</sup>C magnetic resonance spectroscopy demonstrated lower rates of postprandial and insulin-regulated hepatic glycogen synthesis in people with T2D<sup>38,39</sup>. The higher half-maximal effective concentration and lower maximum effect of insulin on hepatic glycogen synthesis<sup>39</sup> indicate impaired IR activation with subsequent posttranslational modifications of the glycogen synthetic machinery and transcriptional regulation of glucokinase (Fig. 1b). Whereas other insulin effects, such as transcriptional DNL activation via sterol receptor enhancing binding protein-1c (SREBP1c), would be expected to be blunted, hepatic insulin resistance is generally associated with increased hepatic TAG and NAFLD. Accordingly, it has been proposed that only the FOXO1-dependent, but not the SREBP1c-dependent branch of insulin signalling, is defective, suggesting selective hepatic insulin resistance<sup>48</sup>. This hypothesis relies on the assumption that DNL is the major source of hepatic TAG and on experiments showing different roles of insulin receptor substrate (IRS)-1 and IRS-2, substrate-specific AKT phosphorylation or intrinsic pathway sensitivities to insulin. Conversely, NEFA re-esterification probably accounts for the majority of hepatic lipogenesis and very low-density lipoprotein (VLDL) secretion<sup>49–51</sup>. Decreased insulin-stimulated hepatic IR kinase activity suggests a common proximal abnormality in T2D<sup>52</sup>. Furthermore, DNL upregulation is not dependent exclusively on IR kinase activity, but can also occur through activation of carbohydrate receptor enhancing binding protein (ChREBP)<sup>53</sup>, mTORC1–SREBP1c<sup>54</sup> and fructose-stimulated pathways<sup>55</sup> (Fig. 1b). A recent study found that fatty acid esterification to TAG is mostly dependent on NEFA delivery to the liver and independent of hepatic insulin signalling<sup>16</sup>. This alternative hypothesis also explains the development of NAFLD through increased NEFA flux derived from increased lipolysis by insulin-resistant WAT.

In addition to caloric overload, macronutrients exert specific effects by modulating enteroendocrine secretion and, in turn, pancreatic islet and brain function before reaching the splanchnic bed to directly

ceramides, which may also arise via sphingomyelin and salvage pathways (5). Ceramides activate protein phosphatase (PP2A) and PKC $\zeta$  inhibiting AKT2 phosphorylation. Amino acids (AA) inhibit IRS1 activation via the mTOR–p70S6 serine kinase (S6K) pathway (6). Independently, inherited and acquired factors can lead to abnormal mitochondrial function (7) accelerating accumulation of DAG and, potentially, acylcarnitine (from incomplete  $\beta$ -oxidation) (8). Finally, chronically elevated reactive oxygen species (ROS) can inhibit IRS1 phosphorylation via NF- $\kappa$ B and JNK pathways (9). These effects combine to decrease glucose transport and glycogen synthesis, thereby contributing to postprandial hyperglycaemia. Dotted lines represent regulation, that is, stimulation or inhibition; thicker lines represent increased (flux through) pathways, thinner lines represent decreased (flux through) pathways. CY, cytosol; ER, endoplasmic reticulum; LD, lipid droplet; MT, mitochondria; PM, plasma membrane; G3P, glycerol-3-phosphate; LPA, lysophosphatidic acid; PA, phosphatidic acid; GPAT, acyl-CoA:glycerol-3-phosphate acyltransferase, AGPAT: acyl-CoA:1-acyl-glycerol-3-phosphate acyltransferase; DGAT2; DAG-O-acyltransferase; PAP; phosphatidate phosphatase; MGL, monoacylglycerol lipase.

pool, which allosterically stimulates gluconeogenesis or activates nutrient-sensitive pathways (ChREBP, mTORC and SREBP) to collectively stimulate the transcriptional DNL program. Elevated hepatic acyl-CoA favours production of *sn*-1,2-DAG, sphingolipids and TAG.



In obese humans with NAFLD, the *sn*-1,2-DAG–PKC $\epsilon$  pathway tightly correlates with hepatic insulin resistance<sup>56–60</sup>, whereas ceramide–JUN N-terminal kinase (JNK) correlates more with hepatic oxidative stress and inflammation<sup>58,61,62</sup> (Fig. 1b). In this context, lowering cellular ceramide by ablating dihydroceramide desaturase 1 increased mitochondrial oxygen flux and improved steatosis and glucose metabolism in insulin-resistant mice<sup>63</sup>. Conversely, mitochondrial C16:0 ceramide, generated by overexpression of ceramide synthase 6 (CerS6), interacts with mitochondrial fission factor (MFF) to promote mitochondrial fragmentation, insulin resistance and steatosis<sup>64</sup>. Silencing of MFF prevented CerS6-dependent metabolic abnormalities despite elevated C16:0 ceramide. This suggests that the effects of ceramides on insulin-stimulated glucose metabolism might result indirectly from impaired mitochondrial function with lower fatty acid oxidation, giving rise to other metabolites, for example, *sn*-1,2-DAG or acetyl-CoA, rather than from direct ceramide interference with insulin signalling. Recent studies indicate a critical role of molecular compartmentation of *sn*-1,2-DAGs, specifically in the plasma membrane, in inducing nPKC translocation and insulin resistance. Mice treated with CGI-58 antisense oligonucleotide exhibit elevated hepatic TAG and DAG in lipid droplets, are protected from lipid-induced hepatic insulin resistance and show reductions in plasma membrane DAG and PKC $\epsilon$  translocation<sup>65</sup>.

Alvarez-Hernandez et al. monitored the earliest diet-induced metabolic alterations by examining the effect of a single oral saturated fat load in healthy humans<sup>66</sup>. This study revealed that saturated fat simultaneously induces insulin resistance in liver, skeletal muscle and WAT, and is associated with 70% higher rates of hepatic gluconeogenesis and 20% lower rates of net hepatic glycogenolysis. Similar studies in mice found upregulated expression of toll-like receptor (TLR) and inflammatory pathways, which might contribute to progression of NAFLD, including non-alcoholic steatohepatitis (NASH)<sup>66</sup>. Of note, chronic overfeeding also increased levels of intestine-derived endotoxins promoting TLR4-induced cytokine release by Kupffer cells<sup>67,68</sup>. Other intestinal functions also affect glycaemia and diabetes risk: integrin  $\beta$ 7-knockout mice, which lack natural small-intestinal intraepithelial T lymphocytes, are metabolically hyperactive and resistant to obesity and diabetes<sup>69</sup>. Finally, dietary habits may affect the gut microbiota, modulating intestinal metabolite release and insulin sensitivity<sup>70</sup>. Humans with T2D and NAFLD show distinct metagenomic signatures along with increased branched-chain amino acids<sup>71,72</sup> and decreased short-chain NEFA<sup>73</sup>, which may affect body weight and metabolism.

In summary, overnutrition and WAT dysfunction lead to increased WAT lipolysis, which promotes insulin-independent hepatic lipogenesis resulting in increased ectopic lipid deposition and increased hepatic gluconeogenesis owing to increased increased acetyl-CoA stimulation of pyruvate carboxylase as well as increased glycerol conversion to glucose. This mechanism obviates the previously reported need to invoke selective hepatic insulin resistance to explain the discordance of increased hepatic lipogenesis occurring simultaneously with increased gluconeogenesis<sup>48</sup> (Fig. 1b). This is in line with recent studies showing that weight loss caused by very-low caloric diets rapidly normalizes hepatic steatosis and insulin resistance in liver, but not intramyocellular lipid content or muscle insulin resistance in individuals with T2D<sup>3,11,74</sup>.

## Insulin resistance in skeletal muscle

Studies using <sup>13</sup>C and <sup>31</sup>P magnetic resonance spectroscopy identified impaired insulin-stimulated glycogen synthesis as the major factor responsible for insulin resistance in muscle and reduced insulin-stimulated glucose transport activity as the rate-controlling step that underlies lower glycogen synthesis in patients with T2D and their insulin-resistant relatives<sup>75–79</sup> (Fig. 2a). Reduced insulin-stimulated glucose transport can be mainly attributed to defective insulin signalling at the level of IR and IRS-1-associated PI3K, which has been observed in one study to occur without altered AKT phosphorylation<sup>80</sup>. Whereas

the majority of studies in humans point to proximal defects in insulin signalling, some experimental models provide evidence for distal abnormalities<sup>81,82</sup>. Glycogen synthase can also be stimulated via insulin regulation of glycogen synthase kinase-3 (GSK3) or independently via allosteric activation by glucose-6-phosphate<sup>83</sup> in skeletal muscle<sup>75–77,84</sup> and liver<sup>38,85</sup>, but its activity does not appear to regulate insulin-stimulated glucose disposal (Figs. 1, 2).

Lean first-degree relatives of patients with T2D present with predominantly muscle insulin resistance<sup>76</sup>. Ingestion of two high-carbohydrate meals revealed their early metabolic abnormalities: ingested carbohydrates were diverted from muscle glycogen synthesis to the liver, where augmented carbohydrate availability and compensatory hyperinsulinaemia promoted hepatic DNL, hepatic TAG synthesis and VLDL secretion, hypertriglyceridaemia and reduced plasma high-density lipoproteins<sup>30</sup>. The critical importance of skeletal muscle is illustrated by the observation that a single bout of exercise, which activates AMP kinase, promotes translocation of the glucose transporter GLUT4 and glucose uptake independently of insulin<sup>86</sup>, completely reversed these abnormalities<sup>76,87</sup>. Insulin-resistant individuals also exhibit reduced muscle mitochondrial density, gene expression and function, which impedes lipid oxidation. This, combined with augmented hepatic TAG release, contributes to muscle lipid accumulation. Collectively, these findings suggest a specific phenotype, whereby genetic and/or acquired reductions in muscle mitochondrial function predispose these individuals to *sn*-1,2-DAG accumulation, activation of PKC $\theta$  and PKC $\epsilon$  and insulin resistance in muscle, which can be enhanced further by excessive production of reactive oxygen species<sup>88</sup> (Fig. 2b). Such selective muscle insulin resistance also increases cardiometabolic risk owing to increased TAG and VLDL production and subsequent dyslipidaemia. The association between muscle insulin resistance and abnormal mitochondrial function represents a frequently observed feature of the elderly and people prone to or with overt T2D<sup>79,89</sup>.

There is increasing evidence supporting a hypothesis whereby gene variants in mitochondrial DNA (mtDNA) and mitochondrial-function-related nuclear DNA contribute to insulin resistance and T2D<sup>90</sup> or abnormal exercise-induced responses<sup>91</sup>. Gene variants in mitochondrial-function-related nuclear DNA lead to relatively mildly impaired mitochondrial function, whereas classical mtDNA gene variants typically cause a severe reduction in mitochondrial function with neurological deficits and  $\beta$ -cell failure. In contrast to genetic and acquired alterations that lead to mild impairments in mitochondrial activity and a predisposition to ectopic lipid accumulation and insulin resistance, alterations that lead to severe reductions in mitochondrial activity (for example, mtDNA variants) result in increased dependency on anaerobic glycolysis, hyperlactaemia and increased glucose metabolism<sup>92–94</sup>. In support of this hypothesis, a recent European GWAS reported that a nonsynonymous variant of *N*-acetyltransferase 2 (*NAT2*) is associated with insulin resistance and related traits as well as with decreased adipocyte differentiation, insulin-mediated glucose uptake and increased WAT lipolysis<sup>95</sup>. Silencing or knocking down the mouse *NAT2* orthologue, *NAT1*, induces insulin resistance, glucose intolerance and exercise intolerance<sup>96,97</sup>, and is associated with ectopic accumulation of TAG and DAG accumulation and activation of hepatic PKC $\epsilon$  and muscle PKC $\theta$ <sup>97</sup>. *Nat1*<sup>−/−</sup> mice also display mild reductions in mitochondrial function and altered morphology, demonstrating another genetic link between reduced mitochondrial function, TAG and DAG deposition and nPKC-induced liver and muscle insulin resistance<sup>97</sup>. Further supporting a relevant role of mitochondria for the development of insulin resistance, mice with muscle-specific knockout of sarcolipin, which is required for mitochondrial sarcoendoplasmic reticulum Ca<sup>2+</sup>-ATPase (SERCA) uncoupling and lipid oxidation, develop obesity and DAG–nPKC-mediated muscle insulin resistance, whereas sarcolipin overexpression prevents obesity-induced insulin resistance<sup>98</sup>. Other gene variants may also predispose humans to muscle insulin resistance and T2D independently from altered mitochondrial function, such as

the AKT2 partial loss-of-function mutation that results in lower insulin-stimulated muscle and adipose glucose uptake<sup>99</sup>, whereas an activating mutation causes fasting hypoglycaemia<sup>100</sup>. *AS160* (also known as *TBC1D4*) gene variants suggest links between insulin signalling and glucose transport leading to muscle insulin resistance, postprandial hyperinsulinaemia and hyperglycaemia<sup>101,102</sup>. Furthermore, RAC1-mediated glucose transport can become dysregulated in insulin-resistant murine and human skeletal muscle<sup>103</sup> (Fig. 2a,b). These genotypes may have been advantageous for preserving glucose for other tissues in the prehistoric arctic environment.

## Adipose tissue dysfunction

Similar to skeletal muscle, insulin-resistant humans exhibit reductions in membrane IR content, IR tyrosine kinase activity and insulin-stimulated glucose uptake in adipose tissue<sup>104–107</sup>. Although WAT accounts for less than 5% of postprandial glucose disposal, it has a disproportionate role in regulating whole-body glucose metabolism through its ability to alter rates of hepatic gluconeogenesis through NEFA and glycerol release<sup>10,14,104</sup>. Furthermore, increased WAT glucose uptake can enhance ChREBP-mediated lipogenesis, provide glycerol-3-phosphate for NEFA esterification or serve as signal for adipokines<sup>108</sup>. Other pathways, such as insulin-regulated  $\beta$ -adrenergic lipolysis or cytokine interaction, may also contribute to dysregulated WAT metabolism<sup>31,32,109</sup>. Insulin resistance in WAT shows temporal differences in the direction of net NEFA flux across the capillary bed between fasting and postprandial states, indicating lower fluctuations in obese men with impaired net adipose fat storage<sup>110</sup>. Finally, compartment-specific differences with higher WAT lipolysis and lower lipogenesis in visceral versus subcutaneous adipose tissue could enhance portal lipid delivery to the liver and contribute to metabolic dysregulation<sup>107</sup>.

Despite the association between body fat mass and insulin resistance, there is accumulating evidence that abnormal adipocyte function as well as liver and muscle lipid metabolites (including *sn*-1,2-DAG and ceramide), but not fat mass per se underlie common insulin resistance. In addition to lifestyle, GWAS studies suggest that genetic variants may affect the association of body fat mass or ectopic fat distribution with glycaemia and insulin resistance<sup>111,112</sup>. Some gene loci recently identified to associate with insulin resistance are associated with insulin resistance at a given level of adiposity and modulate insulin sensitivity via adipocyte differentiation<sup>113</sup>, supporting the concept that limited WAT storage capacity rather than overall fat mass is the main contributor to insulin resistance and associated diseases. Epigenetic adipose tissue modifications may further influence these interactions<sup>114</sup>.

Enlarged WAT mass and adipocyte size has been linked with inadequate vascularization, hypoxia, fibrosis and/or macrophage infiltration with low-grade inflammation<sup>108</sup> (Fig. 1b). High-fat diet and obesity may activate saturated fatty acid-stimulated adenine nucleotide translocase 2 (ANT2), an inner mitochondrial membrane protein, which results in relative adipocyte hypoxia and triggers the transcription factor hypoxia-inducible factor-1 $\alpha$  (HIF-1 $\alpha$ ), setting off adipose dysfunction and inflammation<sup>115</sup>. Adipocyte-specific *Ant2* (also known as *Slc25a5*) deletion improves obesity-induced adipocyte hypoxia by lowering oxygen demand—despite unchanged mitochondrial mass—and, in turn, inflammation and insulin resistance<sup>116</sup>. This suggests that fatty acid-mediated ANT2 activation may be an early event in adipocyte dysfunction and a possible target for novel insulin sensitizers or anti-obesity drugs. Other early events comprise mechanical stress on membranes and extracellular matrix, causing dynamic adaptation until adipocyte death, apoptosis or de-differentiation<sup>108</sup>. During this process, release of chemotactic signals attracts bone marrow-derived pro-inflammatory M1 macrophages, leading to adipose remodelling by a wide range of activities including PPAR $\gamma$ -driven lipid storage, extracellular matrix modification, lysosomal clearance of dead adipocytes and cytokine release (Fig. 3). Compared with acute clinical inflammation, metabolic

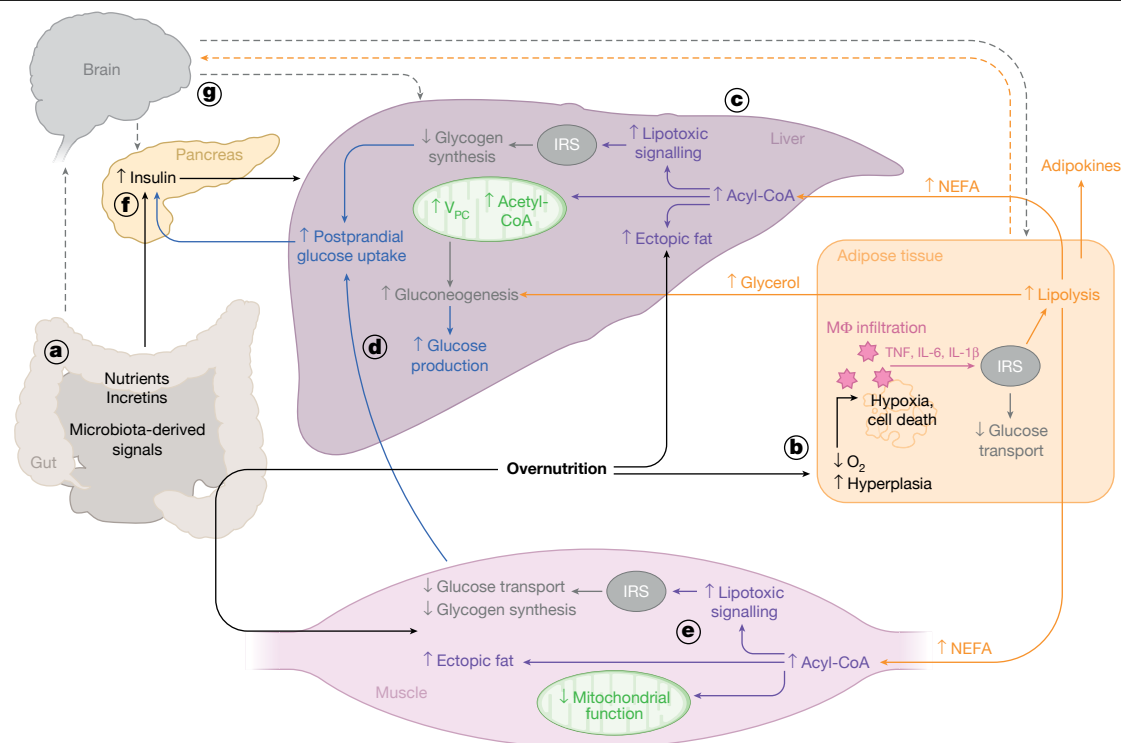
macrophage activation exhibits distinct activation modes<sup>117</sup>, involving chemokine monocyte chemotactic protein 1 (MCP-1) and its receptor CCR2<sup>118</sup>, B2 lymphocytes<sup>119</sup> and/or interferon- $\gamma$  and tumour necrosis factor (TNF), produced by natural killer cells in visceral adipose tissue<sup>120,121</sup>. Local cytokine release (TNF, interleukins (IL)-1 $\beta$  and IL-6) within WAT can suffice to induce adipose insulin resistance without overflow-related effects in distant tissues, usually associated with so-called subclinical or low-grade inflammation. In line with these results, insulin-resistant obese adolescents displayed more than 20-fold higher IL-6 levels in adipose tissue than in plasma, and increased adipose tissue expression of CGI-58 protein, similar to observations in high-fat fed rats<sup>14</sup>. Of course, continued adipose tissue enlargement and concomitant stress will lead to cytokine overflow, creating an imbalance between insulin-sensitizing (including adiponectin and leptin) and pro-inflammatory adipokines (including RBP4, resistin, IL-6 and TNF). These endocrine effects have been intensively studied over the past decades, demonstrating inhibition of IR kinase activity or activation of JNK, oxidative and ER stress. However, the concentrations required to achieve inhibitory effects are often orders of magnitude higher than those measured in plasma from patients with insulin resistance, and pharmacological agents are not generally specific for inflammatory pathways. Recent clinical trials examining IL-1 $\beta$  inhibition with canakinumab, despite causing large reductions in C-reactive protein and IL-6, did not reduce incidence of chronic hyperglycaemia in T2D<sup>122</sup>. Inhibiting the obesity-related kinases, IKK $\beta$  and TBK1, with amlexanox slightly reduced glycaemia, albeit with a paradoxical increase in serum IL-6<sup>123</sup>. Of note, insulin resistance can exist without any relevant adipose tissue inflammation; this is supported by lipodystrophy models in human and mice<sup>95</sup>. Moreover, rats rapidly develop WAT insulin resistance after just several days of high-fat feeding associated with liver and muscle lipid deposition<sup>124</sup>, whereas adipocyte death and macrophage infiltration are detectable only after four weeks<sup>125,126</sup>. Similarly, healthy humans exhibit adipose tissue insulin resistance within hours after saturated fat loading without alteration of circulating anti- or pro-inflammatory markers<sup>66</sup>.

Together, these findings indicate that metabolic changes leading to ectopic lipid accumulation are relatively early events in the pathogenesis of insulin resistance and T2D, whereas WAT inflammation with cytokine spillover represent chronic alterations that occur later, promoting progression to fasting and postprandial hyperglycaemia in conjunction with reduced  $\beta$ -cell function. Recent studies have also implicated other adipose-derived factors, for example, circulating exosomal miRNAs, which may contribute to gene expression in distant tissues and glucose tolerance, as demonstrated in mice lacking the miRNA-processing enzyme Dicer and in lipodystrophic humans<sup>127</sup>.

## Cerebral regulation of hepatic metabolism

High energy requirements and limited energy storage capacity in the brain may explain why cerebral energy supply by glucose and ketones is completely dependent on the liver and, to some extent, kidney, during starvation and nearly independent of direct endocrine regulation. Conversely, cerebral insulin action may affect appetite control, mood, cognitive function and possibly peripheral glucose metabolism<sup>128</sup>. In mice, insulin and leptin act directly on the hypothalamic arcuate nucleus to activate proopiomelanocortin and inhibit Agouti-related protein neurons, whereas adipostatic signals stimulate melanocortin 4-expressing paraventricular neurons to induce satiety and energy expenditure<sup>129</sup>. Hypothalamic inflammation, reflected by higher mediobasal hypothalamic gliosis in obese rodents and humans<sup>130</sup>, has been suggested to lead to chronic central insulin and leptin resistance, which would promote excessive food intake and bodyweight gain. In rodents, central insulin action lowered EGP, hepatic gluconeogenesis, WAT lipolysis and glucagon secretion, but increased muscle glucose uptake<sup>131–135</sup> (Fig. 3). However, carefully controlled studies failed to confirm similar brain insulin action to regulate hepatic glucose fluxes





**Fig. 3 | A unified concept of insulin resistance in humans.** **a**, Overnutrition leads to adipose tissue hypertrophy and hyperplasia and ectopic TAG deposition, mainly in muscle and liver—key features of insulin resistance. **b**, Adipose dysfunction, possibly due to local hypoxia ultimately resulting in apoptosis and cell death, recruits and transforms macrophages to release, for example, TNF and interleukins (IL-1 $\beta$  and IL-6). Local inflammatory reaction increases lipolysis directly or via inhibiting insulin signalling with subsequent release of NEFA and glycerol from WAT. Chronically, adipose dysfunction alters adipocytokine secretion favouring systemic low-grade inflammation. **c**, In liver, glycerol as substrate and NEFA-derived acetyl-CoA, allosterically activating pyruvate carboxylase flux ( $V_{PC}$ ), stimulate gluconeogenesis and fasting glucose production. NEFA and glycerol, as substrates of TAG accumulation, initiate NAFLD in the absence of adequate mitochondrial function and generate lipotoxic metabolites that inhibit insulin signalling. **d**, Hepatocellular insulin resistance not only chronically upregulates

gluconeogenesis, but also decreases insulin-stimulated net glycogen synthesis and glucose uptake, in turn raising postprandial glucose production. **e**, In muscle, increased NEFA availability, accelerated by—possibly inherited— inadequate mitochondrial fat oxidation, also favours lipid synthesis, inhibiting insulin-stimulated glucose transport and glycogen synthesis. This, combined with lower non-insulin-mediated glucose uptake due to sedentary lifestyle, contributes to the postprandial glucose rise. **f**, These different mechanisms, along with direct stimulation by nutrients and enteroendocrine signals (such as GLP-1 and GIP) increase the insulin:glucagon secretion ratio, resulting in normoglycaemia at the expense of hyperinsulinaemia. **g**, Chronically, both acquired and inherited factors impair insulin secretion, with subsequent postprandial and fasting hyperglycaemia. The brain may also contribute to regulation of peripheral metabolism via afferent (for example, leptin from WAT (orange dashed line)) and efferent (for example, to liver (grey dashed line)) signalling.

in awake dogs<sup>136,137</sup>. In humans, intranasal insulin application did not affect fasting EGP, slightly decreased hepatic fat and increased ATP content in glucose-tolerant individuals, but not in people with T2D<sup>138</sup>. Similarly,  $K_{ATP}$ -channel activation decreased EGP only in glucose-tolerant humans<sup>139,140</sup>. Some studies suggested that cerebral insulin action results in parasympathomimetic IL-6 secretion by Kupffer cells to inhibit hepatic gluconeogenesis<sup>133,141</sup>. All of these studies are limited by experimental conditions such as application and dosing of insulin, spillover of intranasally delivered insulin into systemic circulation and suitable metabolic control. Nevertheless, the brain can be involved in other aspects of interorgan crosstalk, orchestrated by metabolites<sup>11</sup>, adipokines (leptin) or enteroendocrine circuits (such as glucagon-like peptide 1, gastric inhibitor peptide, ghrelin, cholecystokinin or fibroblast growth factor (FGF)-19). The hypoleptinaemia-mediated stimulation of the HPA axis with subsequent stimulation of WAT lipolysis<sup>10</sup> might be an example of how the human brain could indirectly regulate hepatic gluconeogenesis and EGP during starvation<sup>12</sup>.

### A unifying concept of the development of T2D

Recent studies assessing rates of hepatic pyruvate carboxylase flux, palmitate turnover and hepatic acetyl-CoA content in an awake rat model

of T2D, revealed a key mechanism by which insulin acutely suppresses hepatic gluconeogenesis and how increased WAT lipolysis, owing to macrophage infiltration with localized inflammation, can increase rates of hepatic gluconeogenesis and cause fasting hyperglycaemia<sup>14</sup>. During hyperinsulinaemic–normoglycaemic clamps, nondiabetic rats exhibited a sequence of events starting with a 90% fall in circulating NEFA and glycerol within 5 min, followed by a 50% reduction in hepatic acetyl-CoA and 70% suppression of EGP within 10 min without affecting hepatic glycogen, circulating lactate or glucagon concentrations. Furthermore, insulin-mediated suppression of WAT lipolysis, leading to reductions in hepatic acetyl-CoA could entirely explain acute insulin-induced suppression of hepatic gluconeogenesis. In line with these results, rodents lacking canonical hepatocellular insulin signalling (AKT1-, AKT2-, FOXO1- or IR-antisense oligonucleotide treatment) showed intact insulin-mediated EGP suppression<sup>14,45,47,48</sup>. Together, these studies demonstrate that—in contrast to the effects of insulin to stimulate hepatic glycogen synthesis through direct stimulation of hepatic insulin signalling—insulin acutely suppresses hepatic gluconeogenesis, mostly via an indirect mechanism through suppression of WAT lipolysis (Fig. 3).

Conversely, rats on a four-week high-fat diet developed fasting hyperglycaemia along with 25% higher rates of EGP, owing to increases in

hepatic pyruvate carboxylase flux and glycerol-to-glucose conversion<sup>14</sup>. During clamps, impaired EGP suppression, along with higher rates of pyruvate carboxylase flux, could be attributed to increased hepatic acetyl-CoA content resulting from greater WAT lipolysis. Inhibition of lipolysis in atglutatin-treated rat models of T2D or high-fat fed adipose triglyceride lipase (ATGL)-knockout mice reversed these abnormalities. Furthermore, this study depicted distinct time-dependent alterations in WAT biology, starting with adipocyte hypertrophy, followed by increased levels of macrophage-secreted granulocyte-macrophage colony-stimulating factor (GM-CSF) and IL-6 in plasma and WAT. Consistent with the potential role for localized macrophage-induced WAT lipolysis, IL-6 infusion stimulated, whereas anti-IL-6 treatment or macrophage-specific JNK knockout ameliorated WAT lipolysis, hepatic insulin resistance and gluconeogenesis. Translating these studies to humans, obese insulin-resistant adolescents also exhibited increased fasting EGP, impaired insulin-mediated suppression of lipolysis, EGP and macrophage infiltration, and 50% higher IL-6 concentrations in WAT<sup>14</sup>. Taken together, these studies indicate that macrophage-induced cytokine-mediated WAT lipolysis raises hepatic acetyl-CoA content and pyruvate carboxylase activity and flux, probably serving as a molecular mechanism linking WAT inflammation to both fasting and postprandial hyperglycaemia (Fig. 3). These data also challenge the canonical view of inflammation-mediated hepatic insulin resistance occurring through activation of the NF- $\kappa$ B–JNK–ceramide biosynthetic pathways and explain the relatively mild metabolic phenotype in rodents with abrogated hepatic insulin action.

Nevertheless, hepatic insulin resistance, concomitant increases in EGP and hyperglycaemia in T2D are probably multifactorial in nature and not exclusively due to increased WAT lipolysis. Supporting this view, a three-day very low-calorie diet reversed hyperglycaemia in a rat model of uncontrolled T2D, not only via reductions in hepatic acetyl-CoA with lower rates of hepatic gluconeogenesis, but also via reductions of hepatic DAG–PKC $\epsilon$ -mediated hepatic insulin resistance and lower rates of hepatic glycogenolysis<sup>11</sup>. Of note, these effects occurred independently of any changes in hepatic ceramides, cytokines, plasma branched-chain amino acids, glucagon, corticosterone or FGF-21.

Holistically, adaptation of metabolic fluxes during fasting and obesity-related diabetes represents the response of WAT to altered substrate supply, which would prevent distant insulin-dependent tissues from substrate oversupply and provide sufficient vital substrates to brain. We postulate that the biology of fasting and postprandial hyperglycaemia depends on dysregulated WAT lipolysis (and possibly contributions from intrahepatic lipolysis) driving hepatic gluconeogenesis through allosteric hepatic acetyl-CoA activation and altered substrate signalling, preferably via the *sn*-1,2-DAG–nPKC pathway (Fig. 3). This concept also highlights important targets for future T2D treatment.

## Outlook

The idea that metabolic flux adaptation in liver and, to some extent, skeletal muscle is largely orchestrated by WAT in health and disease is supported by a series of studies in humans and model organisms. Nevertheless, several aspects still require confirmation both on a molecular–cellular level and in humans under specific metabolic conditions. First, the initial events leading to adipocyte dysfunction and the factors responsible for ectopic lipid deposition in skeletal muscle are not fully understood in humans. Second, the subcellular distribution of different lipid mediators in various compartments and their interactions with nPKC activation and other downstream factors require additional translational studies. In this context, certain intracellular lipids may be linked with insulin sensitivity to varying degrees in sedentary and strongly in physically active humans<sup>142</sup>. Moreover, genetically modified animals such as adipose- and liver-specific PKC $\epsilon$ -knockout mice<sup>143</sup>, are helpful for exploring cellular pathways, but require detailed analysis of the experimental conditions and ultimately testing of their

relevance in humans. Third, the rapidly growing body of multi-omics data might contribute to a better understanding of the cooperative action of metabolites to modify flux rates and insulin signalling. Along these lines, the relevance of metagenomics and epigenomics for the initiation, amplification or reversal of insulin resistance in humans is still largely unclear, despite recently gained insights into the dynamic regulation of insulin sensitivity following metabolic surgery<sup>144</sup>. Recent detection of different T2D phenotypes<sup>35–37</sup> will reinforce investigation of gene variants, metabolites and neuro-immune-endocrine signals for interorgan communication regulating insulin sensitivity<sup>145</sup>. We anticipate that future studies will yield the mechanisms that underlie insulin resistance and  $\beta$ -cell dysfunction, which will guide precision medicine towards more effective treatments for T2D and related disorders such as NAFLD, including NASH and the metabolic syndrome.

1. Cho, N. H. et al. IDF Diabetes Atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res. Clin. Pract.* **138**, 271–281 (2018).
2. Mahajan, A. et al. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat. Genet.* **50**, 559–571 (2018).
3. Petersen, K. F. et al. Reversal of nonalcoholic hepatic steatosis, hepatic insulin resistance, and hyperglycemia by moderate weight reduction in patients with type 2 diabetes. *Diabetes* **54**, 603–608 (2005).
4. Lean, M. E. J. et al. Durability of a primary care-led weight-management intervention for remission of type 2 diabetes: 2-year results of the DIRECT open-label, cluster-randomised trial. *Lancet Diabetes Endocrinol.* **7**, 344–355 (2019).
5. Bellou, V., Belbasis, L., Tzoulaki, I. & Evangelou, E. Risk factors for type 2 diabetes mellitus: an exposure-wide umbrella review of meta-analyses. *PLoS One* **13**, e0194127 (2018).
6. Barrès, R. & Zierath, J. R. The role of diet and exercise in the transgenerational epigenetic landscape of T2DM. *Nat. Rev. Endocrinol.* **12**, 441–451 (2016).
7. Wahl, S. et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541**, 81–86 (2017).
8. Cahill, G. F. Jr. Starvation in man. *N. Engl. J. Med.* **282**, 668–675 (1970).
9. Rothman, D. L., Magnusson, I., Katz, L. D., Shulman, R. G. & Shulman, G. I. Quantitation of hepatic glycogenolysis and gluconeogenesis in fasting humans with <sup>13</sup>C NMR. *Science* **254**, 573–576 (1991).
10. Perry, R. J. et al. Leptin mediates a glucose-fatty acid cycle to maintain glucose homeostasis in starvation. *Cell* **172**, 234–248 (2018).
- This study describes the physiological response to starving, which integrates hormone secretion and metabolic fluxes to promote the shift from carbohydrate to fat metabolism for maintaining gluconeogenesis, reflecting key features of insulin resistance.**
11. Perry, R. J. et al. Mechanisms by which a very-low-calorie diet reverses hyperglycemia in a rat model of type 2 diabetes. *Cell Metab.* **27**, 210–217 (2018).
12. Steinhilber, M. L. et al. The circulating metabolome of human starvation. *JCI Insight* **3**, e121434 (2018).
13. Fazeli, P. K. et al. FGF21 and the late adaptive response to starvation in humans. *J. Clin. Invest.* **125**, 4601–4611 (2015).
14. Perry, R. J. et al. Hepatic acetyl CoA links adipose tissue inflammation to hepatic insulin resistance and type 2 diabetes. *Cell* **160**, 745–758 (2015).
15. Roden, M. et al. Effects of free fatty acid elevation on postabsorptive endogenous glucose production and gluconeogenesis in humans. *Diabetes* **49**, 701–707 (2000).
16. Vatner, D. F. et al. Insulin-independent regulation of hepatic triglyceride synthesis by fatty acids. *Proc. Natl Acad. Sci. USA* **112**, 1143–1148 (2015).
17. Petersen, M. C. et al. Insulin receptor Thr1160 phosphorylation mediates lipid-induced hepatic insulin resistance. *J. Clin. Invest.* **126**, 4361–4371 (2016).
18. Peterson, K. F., Dufour, F., Cline, G. W. & Shulman, G. I. Regulation of hepatic mitochondrial oxidation by glucose-alanine cycling during starvation in humans. *J. Clin. Invest.* **129**, 4671–4675 (2019).
19. Sarabhai, T. & Roden, M. Hungry for your alanine: when liver depends on muscle proteolysis. *J. Clin. Invest.* **129**, 4563–4566 (2019).
20. Pasiakos, S. M., Caruso, C. M., Kellogg, M. D., Kramer, F. M. & Lieberman, H. R. Appetite and endocrine regulators of energy balance after 2 days of energy restriction: insulin, leptin, ghrelin, and DHEA-S. *Obesity* **19**, 1124–1130 (2011).
21. Schorr, M. & Miller, K. K. The endocrine manifestations of anorexia nervosa: mechanisms and management. *Nat. Rev. Endocrinol.* **13**, 174–186 (2017).
22. Ravussin, Y., Leibel, R. L. & Ferrante, A. W. Jr. A missing link in body weight homeostasis: the catabolic signal of the overfed state. *Cell Metab.* **20**, 565–572 (2014).
23. Friedman, J. The long road to leptin. *J. Clin. Invest.* **126**, 4727–4734 (2016).
24. Riddle, M. R. et al. Insulin resistance in cavefish as an adaptation to a nutrient-limited environment. *Nature* **555**, 647–651 (2018).
- Certain cave-adapted fish populations develop diminished insulin signalling in a nutrient-restricted environment, which protects them from blood glucose decline, reflecting a beneficial effect of insulin resistance.**
25. Carrera, P. et al. Substitution of Leu for Pro-193 in the insulin receptor in a patient with a genetic form of severe insulin resistance. *Hum. Mol. Genet.* **2**, 1437–1441 (1993).
26. Abdul-Ghani, M. A. & DeFronzo, R. A. Plasma glucose concentration and prediction of future risk of type 2 diabetes. *Diabetes Care* **32**, S194–S198 (2009).
27. Tabák, A. G. et al. Trajectories of glycaemia, insulin sensitivity, and insulin secretion before diagnosis of type 2 diabetes: an analysis from the Whitehall II study. *Lancet* **373**, 2215–2221 (2009).



28. Ohn, J. H. et al. 10-year trajectory of  $\beta$ -cell function and insulin sensitivity in the development of type 2 diabetes: a community-based prospective cohort study. *Lancet Diabetes Endocrinol.* **4**, 27–34 (2016).
  29. DeFronzo, R. A. & Tripathy, D. Skeletal muscle insulin resistance is the primary defect in type 2 diabetes. *Diabetes Care* **32**, S157–S163 (2009).
  30. Petersen, K. F. et al. The role of skeletal muscle insulin resistance in the pathogenesis of the metabolic syndrome. *Proc. Natl Acad. Sci. USA* **104**, 12587–12594 (2007).
  31. Saltiel, A. R. & Olefsky, J. M. Inflammatory mechanisms linking obesity and metabolic disease. *J. Clin. Invest.* **127**, 1–4 (2017).
  32. Guilherme, A., Henriques, F., Bedard, A. H. & Czech, M. P. Molecular pathways linking adipose innervation to insulin action in obesity and diabetes mellitus. *Nat. Rev. Endocrinol.* **15**, 207–225 (2019).
  33. Umpierrez, G. E., Smiley, D. & Kitabchi, A. E. Narrative review: ketosis-prone type 2 diabetes mellitus. *Ann. Intern. Med.* **144**, 350–357 (2006).
  34. Petersen, K. F. et al. Increased prevalence of insulin resistance and nonalcoholic fatty liver disease in Asian-Indian men. *Proc. Natl Acad. Sci. USA* **103**, 18273–18277 (2006).
  35. Zaharia, O. P. et al. Clusters of patients with recent-onset diabetes show different risk profiles for diabetes-associated diseases during a 5-year follow-up. *Lancet. Diabetol. Endocrinol.* **7**, 684–694 (2019).
- A subgroup of people with diabetes exhibit severe insulin resistance along with higher ectopic fat accumulation and increased risk of comorbidities, which require specific attention for precise prevention and treatment.**
36. Ahlqvist, E. et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol.* **6**, 361–369 (2018).
  37. Udler, M. S. et al. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: a soft clustering analysis. *PLoS Med.* **15**, e1002654 (2018).
  38. Magnusson, I., Rothman, D. L., Katz, L. D., Shulman, R. G. & Shulman, G. I. Increased rate of gluconeogenesis in type II diabetes mellitus. A  $^{13}\text{C}$  nuclear magnetic resonance study. *J. Clin. Invest.* **90**, 1323–1327 (1992).
  39. Krssak, M. et al. Alterations in postprandial hepatic glycogen metabolism in type 2 diabetes. *Diabetes* **53**, 3048–3056 (2004).
  40. Rizza, R. A. Pathogenesis of fasting and postprandial hyperglycemia in type 2 diabetes: implications for therapy. *Diabetes* **59**, 2697–2707 (2010).
  41. Gastaldello, A. et al. Influence of obesity and type 2 diabetes on gluconeogenesis and glucose output in humans: a quantitative study. *Diabetes* **49**, 1367–1373 (2000).
  42. Rebrin, K., Steil, G. M., Mittelman, S. D. & Bergman, R. N. Causal linkage between insulin suppression of lipolysis and suppression of liver glucose output in dogs. *J. Clin. Invest.* **98**, 741–749 (1996).
  43. Buettner, C. et al. Severe impairment in liver insulin signaling fails to alter hepatic insulin action in conscious mice. *J. Clin. Invest.* **115**, 1306–1313 (2005).
  44. Cherrington, A. D. The role of hepatic insulin receptors in the regulation of glucose production. *J. Clin. Invest.* **115**, 1136–1139 (2005).
  45. Lu, M. et al. Insulin regulates liver metabolism in vivo in the absence of hepatic Akt and Foxo1. *Nat. Med.* **18**, 388–395 (2012).
  46. O'Sullivan, I. et al. FoxO1 integrates direct and indirect effects of insulin on hepatic glucose production and glucose utilization. *Nat. Commun.* **6**, 7079 (2015).
  47. Titchenell, P. M., Chu, Q., Monks, B. R. & Birnbaum, M. J. Hepatic insulin signalling is dispensable for suppression of glucose output by insulin in vivo. *Nat. Commun.* **6**, 7078 (2015).
- This mouse study showed that in the absence of FOXO1, insulin signals independently of the hepatic insulin receptor-AKT-FOXO1 axis via an intermediary extrahepatic tissue to regulate hepatic glucose production.**
48. Brown, M. S. & Goldstein, J. L. Selective versus total insulin resistance: a pathogenic paradox. *Cell Metab.* **7**, 95–96 (2008).
  49. Donnelly, K. L. et al. Sources of fatty acids stored in liver and secreted via lipoproteins in patients with nonalcoholic fatty liver disease. *J. Clin. Invest.* **115**, 1343–1351 (2005).
  50. Albert, J. S. et al. Null mutation in hormone-sensitive lipase gene and risk of type 2 diabetes. *N. Engl. J. Med.* **370**, 2307–2315 (2014).
  51. Ali, A. H., Mundi, M., Koutsari, C., Bernlohr, D. A. & Jensen, M. D. Adipose tissue free fatty acid storage in vivo: effects of insulin versus niacin as a control for suppression of lipolysis. *Diabetes* **64**, 2828–2835 (2015).
  52. Caro, J. F. et al. Studies on the mechanism of insulin resistance in the liver from humans with noninsulin-dependent diabetes. Insulin action and binding in isolated hepatocytes, insulin receptor structure, and kinase activity. *J. Clin. Invest.* **78**, 249–258 (1986).
  53. Abdul-Wahed, A., Guilmeau, S. & Postic, C. Sweet sixteenth for ChREBP: established roles and future goals. *Cell Metab.* **26**, 324–341 (2017).
  54. Li, S., Brown, M. S. & Goldstein, J. L. Bifurcation of insulin signaling pathway in rat liver: mTORC1 required for stimulation of lipogenesis, but not inhibition of gluconeogenesis. *Proc. Natl Acad. Sci. USA* **107**, 3441–3446 (2010).
  55. Herman, M. A. & Samuel, V. T. The sweet path to metabolic demise: fructose and lipid synthesis. *Trends Endocrinol. Metab.* **27**, 719–730 (2016).
  56. Kumashiro, N. et al. Cellular mechanism of insulin resistance in nonalcoholic fatty liver disease. *Proc. Natl Acad. Sci. USA* **108**, 16381–16385 (2011).
  57. Magkos, F. et al. Intrahepatic diacylglycerol content is associated with hepatic insulin resistance in obese subjects. *Gastroenterology* **142**, 1444–1446 (2012).
  58. Luukkonen, P. K. et al. Hepatic ceramides dissociate steatosis and insulin resistance in patients with non-alcoholic fatty liver disease. *J. Hepatol.* **64**, 1167–1175 (2016).
  59. ter Horst, K. W. et al. Hepatic diacylglycerol-associated protein kinase  $C_\alpha$  translocation links hepatic steatosis to hepatic insulin resistance in humans. *J. Hepatol.* **64**, 1167–1175 (2016).
  60. Ruby, M. A. et al. Human carboxylesterase 2 reverses obesity-induced diacylglycerol accumulation and glucose intolerance. *Cell Rep.* **18**, 636–646 (2017).
  61. Apostolopoulou, M. et al. Specific hepatic sphingolipids relate to insulin resistance, oxidative stress, and inflammation in nonalcoholic steatohepatitis. *Diabetes Care* **41**, 1235–1243 (2018).
  62. Koliaki, C. et al. Adaptation of hepatic mitochondrial function in humans with non-alcoholic fatty liver is lost in steatohepatitis. *Cell Metab.* **21**, 739–746 (2015).
  63. Chaurasia, B. et al. Targeting a ceramide double bond improves insulin resistance and hepatic steatosis. *Science* **365**, 386–392 (2019).
  64. Hammerschmidt, P. et al. CerS6-derived sphingolipids interact with Mff and promote mitochondrial fragmentation in obesity. *Cell* **177**, 1536–1552 (2019).
  65. Cantley, J. L. et al. CGI-58 knockdown sequesters diacylglycerols in lipid droplets/ER-preventing diacylglycerol-mediated hepatic insulin resistance. *Proc. Natl Acad. Sci. USA* **110**, 1869–1874 (2013).
  66. Hernández, E. Á. et al. Acute dietary fat intake initiates alterations in energy metabolism and insulin resistance. *J. Clin. Invest.* **127**, 695–708 (2017).
  67. Parks, E., Yki-Järvinen, H. & Hawkins, M. Out of the frying pan: dietary saturated fat influences nonalcoholic fatty liver disease. *J. Clin. Invest.* **127**, 454–456 (2017).
  68. Luukkonen, P. K. et al. Saturated fat is more metabolically harmful for the human liver than unsaturated fat or simple sugars. *Diabetes Care* **41**, 1732–1739 (2018).
  69. He, S. et al. Gut intraepithelial T cells calibrate metabolism and accelerate cardiovascular disease. *Nature* **566**, 115–119 (2019).
  70. Ussar, S. et al. Interactions between gut microbiota, host genetics and diet modulate the predisposition to obesity and metabolic syndrome. *Cell Metab.* **22**, 516–530 (2015).
  71. Pedersen, H. K. et al. Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature* **535**, 376–381 (2016).
  72. Hoyle, L. et al. Molecular phenomics and metagenomics of hepatic steatosis in non-diabetic obese women. *Nat. Med.* **24**, 1070–1080 (2018).
  73. Sanna, S. et al. Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat. Genet.* **51**, 600–605 (2019).
  74. Taylor, R. et al. Remission of human type 2 diabetes requires decrease in liver and pancreas fat content but is dependent upon capacity for  $\beta$  cell recovery. *Cell Metab.* **28**, 547–556 (2018).
  75. Cline, G. W. et al. Impaired glucose transport as a cause of decreased insulin-stimulated muscle glycogen synthesis in type 2 diabetes. *N. Engl. J. Med.* **341**, 240–246 (1999).
  76. Perseghin, G. et al. Increased glucose transport-phosphorylation and muscle glycogen synthesis after exercise training in insulin-resistant subjects. *N. Engl. J. Med.* **335**, 1357–1362 (1996).
  77. Roden, M. et al. Mechanism of free fatty acid-induced insulin resistance in humans. *J. Clin. Invest.* **97**, 2859–2865 (1996).
  78. Dresner, A. et al. Effects of free fatty acids on glucose transport and IRS-1-associated phosphatidylinositol 3-kinase activity. *J. Clin. Invest.* **103**, 253–259 (1999).
  79. Szendroedi, J. et al. Muscle mitochondrial ATP synthesis and glucose transport/phosphorylation in type 2 diabetes. *PLoS Med.* **4**, e154 (2007).
  80. Kim, Y. B., Nikoulina, S. E., Ciaraldi, T. P., Henry, R. R. & Kahn, B. B. Normal insulin-dependent activation of Akt/protein kinase B, with diminished activation of phosphoinositide 3-kinase, in muscle in type 2 diabetes. *J. Clin. Invest.* **104**, 733–741 (1999).
  81. Fazakerley, D. J., Krycer, J. R., Kearney, A. L., Hocking, S. L. & James, D. E. Muscle and adipose tissue insulin resistance: malady without mechanism? *J. Lipid Res.* **60**, 1720–1732 (2019).
  82. Czech, M. P. Insulin action and resistance in obesity and type 2 diabetes. *Nat. Med.* **23**, 804–814 (2017).
  83. Wan, M. et al. A noncanonical, GSK3-independent pathway controls postprandial hepatic glycogen deposition. *Cell Metab.* **18**, 99–105 (2013).
  84. Bouskila, M. et al. Allosteric regulation of glycogen synthase controls glycogen synthesis in muscle. *Cell Metab.* **12**, 456–466 (2010).
  85. von Wilamowitz-Moellendorf, A. et al. Glucose-6-phosphate-mediated activation of liver glycogen synthase plays a key role in hepatic glycogen synthesis. *Diabetes* **62**, 4070–4082 (2013).
  86. Musi, N. et al. AMP-activated protein kinase (AMPK) is activated in muscle of subjects with type 2 diabetes during exercise. *Diabetes* **50**, 921–927 (2001).
  87. Rabøl, R., Petersen, K. F., Dufour, S., Flannery, C. & Shulman, G. I. Reversal of muscle insulin resistance with exercise reduces postprandial hepatic de novo lipogenesis in insulin resistant individuals. *Proc. Natl Acad. Sci. USA* **108**, 13705–13709 (2011).
  88. Rueggsegger, G. N., Creio, A. L., Cortes, T. M., Dasari, S. & Nair, K. S. Altered mitochondrial function in insulin-deficient and insulin-resistant states. *J. Clin. Invest.* **128**, 3671–3681 (2018).
  89. Petersen, K. F. et al. Mitochondrial dysfunction in the elderly: possible role in insulin resistance. *Science* **300**, 1140–1142 (2003).
  90. Kraja, A. T. et al. Associations of mitochondrial and nuclear mitochondrial variants and genes with seven metabolic traits. *Am. J. Hum. Genet.* **104**, 112–138 (2019).
  91. Kacerovsky-Bielez, G. et al. Short-term exercise training does not stimulate skeletal muscle ATP synthesis in relatives of humans with type 2 diabetes. *Diabetes* **58**, 1333–1341 (2009).
  92. Holloszy, J. O. “Deficiency” of mitochondria in muscle does not cause insulin resistance. *Diabetes* **62**, 1036–1040 (2013).
  93. Pospisilik, J. A. et al. Targeted deletion of AIF decreases mitochondrial oxidative phosphorylation and protects from obesity and diabetes. *Cell* **131**, 476–491 (2007).
  94. Koh, J. H. et al. TFAM enhances fat oxidation and attenuates high fat diet induced insulin resistance in skeletal muscle. *Diabetes* **68**, 1552–1564 (2019).
  95. Lotta, L. A. et al. Integrative genomic analysis implicates limited peripheral adipose storage capacity in the pathogenesis of human insulin resistance. *Nat. Genet.* **49**, 17–26 (2017).
  96. Knowles, J. W. et al. Identification and validation of *N*-acetyltransferase 2 as an insulin sensitivity gene. *J. Clin. Invest.* **125**, 1739–1751 (2015).
  97. Chennamsetty, I. et al. *Nat1* deficiency is associated with mitochondrial dysfunction and exercise intolerance in mice. *Cell Rep.* **17**, 527–540 (2016).
  98. Maurya, S. K. et al. Sarcoplasmic signaling promotes mitochondrial biogenesis and oxidative metabolism in skeletal muscle. *Cell Rep.* **24**, 2919–2931 (2018).

99. Latva-Rasku, A. et al. A partial loss-of-function variant in AKT2 is associated with reduced insulin-mediated glucose uptake in multiple insulin-sensitive tissues: a genotype-based callback positron emission tomography study. *Diabetes* **67**, 334–342 (2018).
100. Hussain, K. et al. An activating mutation of AKT2 and human hypoglycemia. *Science* **334**, 474 (2011).
101. Dash, S. et al. A truncation mutation in *TBC1D4* in a family with acanthosis nigricans and postprandial hyperinsulinemia. *Proc. Natl Acad. Sci. USA* **106**, 9350–9355 (2009).
102. Moltke, I. et al. A common Greenlandic *TBC1D4* variant confers muscle insulin resistance and type 2 diabetes. *Nature* **512**, 190–193 (2014).
103. Sylow, L. et al. Rac1 signaling is required for insulin-stimulated glucose uptake and is dysregulated in insulin-resistant murine and human skeletal muscle. *Diabetes* **62**, 1865–1875 (2013).
104. Kahn, C. R. Insulin resistance, insulin insensitivity, and insulin unresponsiveness: a necessary distinction. *Metabolism* **27**, 1893–1902 (1978).
105. Freidenberg, G. R., Reichart, D., Olefsky, J. M. & Henry, R. R. Reversibility of defective adipocyte insulin receptor kinase activity in non-insulin-dependent diabetes mellitus. Effect of weight loss. *J. Clin. Invest.* **82**, 1398–1406 (1988).
106. Kahn, B. B. & Flier, J. S. Obesity and insulin resistance. *J. Clin. Invest.* **106**, 473–481 (2000).
107. Bódis, K. & Roden, M. Energy metabolism of white adipose tissue and insulin resistance in humans. *Eur. J. Clin. Invest.* **48**, e13017 (2018).
108. Scherer, P. E. The many secret lives of adipocytes: implications for diabetes. *Diabetologia* **62**, 223–232 (2019).
109. Zeng, X. et al. Innervation of thermogenic adipose tissue via a calyntenin 3β–S100b axis. *Nature* **569**, 229–235 (2019).
110. McQuaid, S. E. et al. Downregulation of adipose tissue fatty acid trafficking in obesity: a driver for ectopic fat deposition? *Diabetes* **60**, 47–55 (2011).
111. Manning, A. K. et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **44**, 659–669 (2012).
112. Shungin, D. et al. New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).
113. Camporez, J. P. et al. Mechanism by which arylamine *N*-acetyltransferase 1 ablation causes insulin resistance in mice. *Proc. Natl Acad. Sci. USA* **114**, E11285–E11292 (2017).
114. Orozco, L. D. et al. Epigenome-wide association in adipose tissue from the METSIM cohort. *Hum. Mol. Genet.* **27**, 1830–1846 (2018).
115. Lee, Y. S. et al. Increased adipocyte O<sub>2</sub> consumption triggers HIF-1α, causing inflammation and insulin resistance in obesity. *Cell* **157**, 1339–1352 (2014).
116. Seo, J. B. et al. Knockdown of ANT2 reduces adipocyte hypoxia and improves insulin resistance in obesity. *Nat. Metab.* **1**, 86–97 (2019).
- Using mouse models, this study shows that adipocyte oxygen demand rather than oxygen supply or angiogenesis is the key determinant of intracellular hypoxia, which may be the initial event leading to adipose tissue inflammation.**
117. Kratz, M. et al. Metabolic dysfunction drives a mechanistically distinct proinflammatory phenotype in adipose tissue macrophages. *Cell Metab.* **20**, 614–625 (2014).
118. Sartipy, P. & Loskutoff, D. J. Monocyte chemoattractant protein 1 in obesity and insulin resistance. *Proc. Natl Acad. Sci. USA* **100**, 7265–7270 (2003).
119. Ying, W. et al. Adipose tissue B2 cells promote insulin resistance through leukotriene LTB4/LTB4R1 signaling. *J. Clin. Invest.* **127**, 1019–1030 (2017).
120. Lee, B.-C. et al. Adipose natural killer cells regulate adipose tissue macrophages to promote insulin resistance in obesity. *Cell Metab.* **23**, 685–698 (2016).
121. Wensveen, F. M. et al. NK cells link obesity-induced adipose stress to inflammation and insulin resistance. *Nat. Immunol.* **16**, 376–385 (2015).
122. Everett, B. M. et al. Anti-inflammatory therapy with canakinumab for the prevention and management of diabetes. *J. Am. Coll. Cardiol.* **71**, 2392–2401 (2018).
123. Oral, E. A. et al. Inhibition of IKKε and TBK1 improves glucose control in a subset of patients with type 2 diabetes. *Cell Metab.* **26**, 157–170 (2017).
124. Samuel, V. T. et al. Inhibition of protein kinase Cε prevents hepatic insulin resistance in nonalcoholic fatty liver disease. *J. Clin. Invest.* **117**, 739–745 (2007).
125. Nishimura, S. et al. CD8<sup>+</sup> effector T cells contribute to macrophage recruitment and adipose tissue inflammation in obesity. *Nat. Med.* **15**, 914–920 (2009).
126. Strissel, K. J. et al. Adipocyte death, adipose tissue remodeling, and obesity complications. *Diabetes* **56**, 2910–2918 (2007).
127. Thomou, T. et al. Adipose-derived circulating miRNAs regulate gene expression in other tissues. *Nature* **542**, 450–455 (2017).
128. Kullmann, S. et al. Brain insulin resistance at the crossroads of metabolic and cognitive disorders in humans. *Physiol. Rev.* **96**, 1169–1209 (2016).
129. Jais, A. & Brüning, J. C. Hypothalamic inflammation in obesity and metabolic disease. *J. Clin. Invest.* **127**, 24–32 (2017).
130. Thaler, J. P. et al. Obesity is associated with hypothalamic injury in rodents and humans. *J. Clin. Invest.* **122**, 153–162 (2012).
131. Obici, S., Zhang, B. B., Karkanias, G. & Rossetti, L. Hypothalamic insulin signaling is required for inhibition of glucose production. *Nat. Med.* **8**, 1376–1382 (2002).
132. Pocai, A. et al. Hypothalamic K<sub>ATP</sub> channels control hepatic glucose production. *Nature* **434**, 1026–1031 (2005).
133. Inoue, H. et al. Role of hepatic STAT3 in brain-insulin action on hepatic glucose production. *Cell Metab.* **3**, 267–275 (2006).
134. Gelling, R. W. et al. Insulin action in the brain contributes to glucose lowering during insulin treatment of diabetes. *Cell Metab.* **3**, 67–73 (2006).
135. Scherer, T. et al. Brain insulin controls adipose tissue lipolysis and lipogenesis. *Cell Metab.* **13**, 183–194 (2011).
136. Ramnanan, C. J., Edgerton, D. S. & Cherrington, A. D. Evidence against a physiologic role for acute changes in CNS insulin action in the rapid regulation of hepatic glucose production. *Cell Metab.* **15**, 656–664 (2012).
- This perspective discusses the evidence that the brain can sense insulin and regulate hepatic glucoregulatory enzyme expression, although the action of cerebral insulin is not essential for the rapid insulin-mediated suppression of glucose production.**
137. Winnick, J. J. et al. Hepatic glycogen can regulate hypoglycemic counterregulation via a liver–brain axis. *J. Clin. Invest.* **126**, 2236–2248 (2016).
138. Gancheva, S. et al. Intranasal insulin lowers hepatic fat accumulation and improves energy metabolism in humans. *Diabetes* **64**, 1966–1975 (2015).
139. Kishore, P. et al. Activation of K<sub>ATP</sub> channels suppresses glucose production in humans. *J. Clin. Invest.* **121**, 4916–4920 (2011).
140. Esterson, Y. B. et al. Central regulation of glucose production may be impaired in type 2 diabetes. *Diabetes* **65**, 2569–2579 (2016).
141. Kimura, K. et al. Central insulin action activates Kupffer cells by suppressing hepatic vagal activation via the nicotinic alpha 7 acetylcholine receptor. *Cell Rep.* **14**, 2362–2374 (2016).
142. Perreault, L. et al. Intracellular localization of diacylglycerols and sphingolipids influences insulin sensitivity and mitochondrial function in human skeletal muscle. *JCI Insight* **3**, e96805 (2018).
143. Brandon, A. E. et al. Protein kinase C epsilon deletion in adipose tissue, but not in liver, improves glucose tolerance. *Cell Metab.* **29**, 183–191 (2019).
144. Gancheva, S. et al. Dynamic changes of muscle insulin sensitivity after metabolic surgery. *Nat. Commun.* **10**, 4179 (2019).
145. Parker, B. L. et al. An integrative systems genetic analysis of mammalian lipid metabolism. *Nature* **567**, 187–193 (2019).

**Acknowledgements** This research is supported by grants from the German Federal Ministry of Health and Ministry of Culture and Science of the state North Rhine-Westphalia to DDZ, the German Federal Ministry of Education and Research to DZD, European Funds for Regional Development (EFRE-0400191), EUREKA Eurostars-2 (E! 113230 DIA-PEP) and the German Science Foundation (CRC/SFB 1116/2 B12) (to M.R.) and by grants from the US Public Health Service (R01 DK-113984, R01 DK114793, R01 DK116774, R01 DK119968, P30 DK-045735) (to G.I.S.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

**Author contributions** M.R. and G.I.S. wrote the manuscript.

**Competing interests** M.R. is on the scientific advisory boards of Bristol-Myers Squibb, Eli Lilly, Gilead Sciences, NovoNordisk, Servier Laboratories, Target Pharmasolutions and Terra Firma and receives investigator-initiated support from Boehringer Ingelheim, Nutricia/Danone and Sanofi-Aventis. G.I.S. is on the scientific advisory boards of Merck, NovoNordisk, Gilead Sciences, AstraZeneca, Aegerion, iMBP, Janssen Research and Development and receives investigator-initiated support from Gilead Sciences, Merck and AstraZeneca.

## Additional information

**Correspondence and requests for materials** should be addressed to M.R. or G.I.S.

**Peer review information** *Nature* thanks Kei Sakamoto and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

© Springer Nature Limited 2019



# Accretion of a giant planet onto a white dwarf star

<https://doi.org/10.1038/s41586-019-1789-8>

Received: 6 June 2019

Accepted: 13 September 2019

Published online: 4 December 2019

Boris T. Gänsicke<sup>1,2\*</sup>, Matthias R. Schreiber<sup>3</sup>, Odette Toloza<sup>1</sup>, Nicola P. Gentile Fusillo<sup>1</sup>, Detlev Koester<sup>4</sup> & Christopher J. Manser<sup>1</sup>

The detection<sup>1</sup> of a dust disk around the white dwarf star G29-38 and transits from debris orbiting the white dwarf WD 1145+017 (ref.<sup>2</sup>) confirmed that the photospheric trace metals found in many white dwarfs<sup>3</sup> arise from the accretion of tidally disrupted planetesimals<sup>4</sup>. The composition of these planetesimals is similar to that of rocky bodies in the inner Solar System<sup>5</sup>. Gravitational scattering of planetesimals towards the white dwarf requires the presence of more massive bodies<sup>6</sup>, yet no planet has so far been detected at a white dwarf. Here we report optical spectroscopy of a hot (about 27,750 kelvin) white dwarf, WD J091405.30+191412.25, that is accreting from a circumstellar gaseous disk composed of hydrogen, oxygen and sulfur at a rate of about  $3.3 \times 10^9$  grams per second. The composition of this disk is unlike all other known planetary debris around white dwarfs<sup>7</sup>, but resembles predictions for the makeup of deeper atmospheric layers of icy giant planets, with H<sub>2</sub>O and H<sub>2</sub>S being major constituents. A giant planet orbiting a hot white dwarf with a semi-major axis of around 15 solar radii will undergo substantial evaporation with expected mass loss rates comparable to the accretion rate that we observe onto the white dwarf. The orbit of the planet is most probably the result of gravitational interactions, indicating the presence of additional planets in the system. We infer an occurrence rate of approximately 1 in 10,000 for spectroscopically detectable giant planets in close orbits around white dwarfs.

WD J091405.30+191412.25 (henceforth WD J0914+1914) was initially classified as a short-period interacting white dwarf binary on the basis of a weak H $\alpha$  emission line detected in its spectrum obtained by the Sloan Digital Sky Survey (SDSS)<sup>8</sup>. Upon closer inspection of this spectrum, we identified additional emission lines of oxygen (O I at wavelengths 7,774 Å and 8,446 Å), and an emission line near 4,068 Å that we tentatively identified as [S II]. The line flux ratios of the hydrogen and oxygen lines are extremely atypical for any white dwarf binary, casting doubt on the published classification.

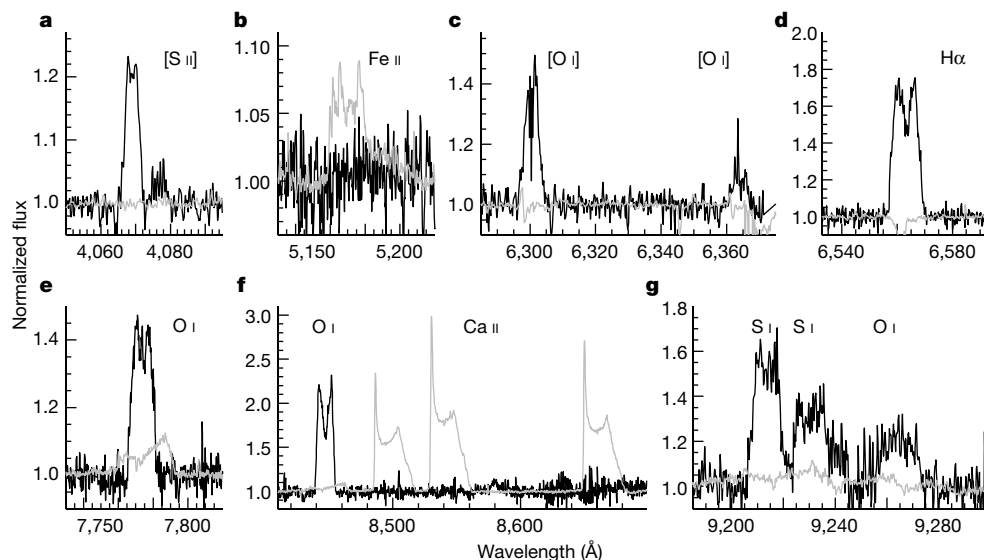
We obtained deep spectroscopy of this star using the X-Shooter spectrograph on the Very Large Telescope of the European Southern Observatory (see Fig. 1), which confirms the presence of [S II] (4,068 Å), and contains additional emission lines of [O I] (6,300 Å and 6,363 Å) as well as a blend of O I and S I lines near 9,200 Å.

The double-peaked morphology of the H $\alpha$  and the O I (8,446 Å) emission lines (see Fig. 1) indicates an origin in a circumstellar gas disk<sup>9</sup>, reminiscent of several white dwarfs with dusty and gaseous planetary debris disks<sup>10,11</sup>. However, the spectra of all known gaseous debris disks are dominated by the emission lines of the Ca II triplet (8,600 Å), with weaker lines of Fe II, which are absent in the X-Shooter observations of WD J0914+1914. Moreover, none of the other gaseous debris disks around white dwarfs show H $\alpha$  emission.

The X-shooter spectrum of WD J0914+1914 displays strong Balmer lines, implying a hydrogen-dominated atmosphere, as well as numerous sharp absorption lines of oxygen and sulfur (see Fig. 2). We determined the white dwarf's effective temperature of  $T_{\text{eff}} = 27,743 \pm 310$  K and a surface gravity of  $\log(g) = 7.85 \pm 0.06$  from the well flux-calibrated SDSS spectra (see Extended Data Fig. 1 and Extended Data Table 1). Fixing these two atmospheric parameters, we measured the photospheric abundances of oxygen and sulfur,  $\log(\text{O}/\text{H}) = -3.25 \pm 0.20$  and  $\log(\text{S}/\text{O}) = -4.15 \pm 0.20$ , and derived upper limits for twelve additional elements (see Fig. 3 and Extended Data Table 2). WD J0914+1914 is accreting at a rate of about  $3.3 \times 10^9 \text{ g s}^{-1}$ , which is among the highest rates observed for hydrogen-atmosphere white dwarfs polluted by planetary debris<sup>3</sup>. However, the measured accretion rate in WD J0914+1914 includes only oxygen and sulfur, and the influxes of these two elements are an order of magnitude larger than in any other of these systems. If thermohaline mixing or convective overshoot are efficient in the atmosphere of WD J0914+1914, the accretion rate could be an order of magnitude higher<sup>12</sup>.

We used the spectral synthesis code Cloudy<sup>13</sup> to model the photoionization of the circumstellar gas disk by the intense ultraviolet flux from the white dwarf (see Methods and Extended Data Figs. 2, 3). The emission lines, which are Doppler-broadened by the Keplerian rotation in the disk<sup>9</sup>, originate from a gaseous disk extending to approximately

<sup>1</sup>Department of Physics, University of Warwick, Coventry, UK. <sup>2</sup>Centre for Exoplanets and Habitability, University of Warwick, Coventry, UK. <sup>3</sup>Institute of Physics and Astronomy, Millennium Nucleus for Planet Formation (NPF), Universidad de Valparaíso, Valparaíso, Chile. <sup>4</sup>Institut für Theoretische Physik und Astrophysik, Universität Kiel, Kiel, Germany. \*e-mail: boris.gaensicke@warwick.ac.uk



**Fig. 1 | Emission lines from the circumstellar disk at WD J0914+1914.** The X-Shooter spectrum of WD J0914+1914 (black) contains strong and broad emission lines of hydrogen, oxygen and sulfur. H $\alpha$  (d) and O I 8,446 Å (f) are double-peaked, indicating an origin in a circumstellar disk<sup>9</sup>. O I 7,774 Å (e) and the oxygen and sulfur lines near 9,240 Å (g) are multiplets, resulting in more complex line profiles. The forbidden sulfur and oxygen lines (a, c) have a smaller peak separation, indicating that they are emitted by material extending

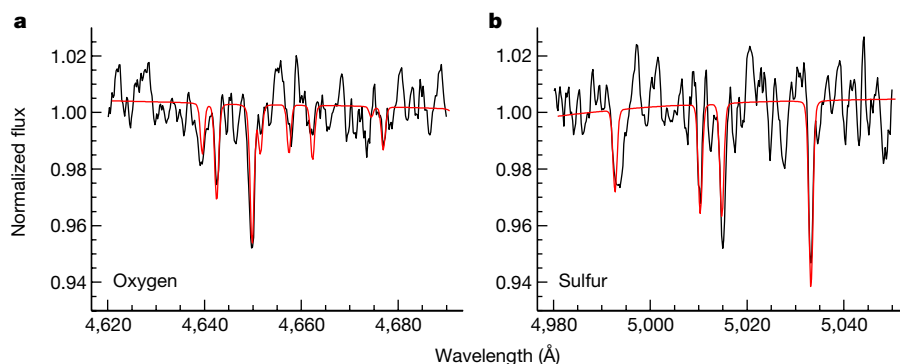
to larger distances from the white dwarf compared to the other lines. The spectra of the gaseous planetary debris disks detected at several other white dwarfs<sup>10,11</sup> are all dominated by the Ca II 8,600 Å triplet (f), with weak additional emission lines of oxygen (e, f) and iron (b), as illustrated by the spectrum of the prototypical system SDSS J1228+1040<sup>28</sup> (grey). The striking difference between the two spectra illustrates the different composition of the planetary material: gaseous in WD J0914+1914 and rocky in SDSS J1228+1040.

1–10 $R_{\odot}$  (where  $R_{\odot}$  is the radius of the Sun; see Extended Data Figs. 2, 4) from the white dwarf, at a density of  $\rho \approx 10^{-11.3}$  g cm<sup>-3</sup>. The relative abundances of oxygen and sulfur derived from this model,  $\log(S/O) = -0.5$ , are consistent with those measured from the photospheric analysis. Hydrogen in the disk is strongly depleted with respect to oxygen and sulfur,  $\log(O/H) = 0.29$  and  $\log(S/H) = -0.21$ . The non-detection of emission lines from other elements apart from hydrogen, oxygen and sulfur allows stringent upper limits to be placed on the abundances of sodium, silicon, calcium and iron in the disk (see Fig. 3 and Extended Data Table 2).

The abundances of the gaseous circumstellar disk, and of the trace metals in the photosphere of WD J0914+1914 are distinctly non-solar and inconsistent with accretion from the wind of a low-mass companion star<sup>14</sup>. A stellar companion is also ruled out by the stringent upper limits on the radial velocity variations of the white dwarf and the absence of an infrared excess (see Methods). In contrast to the white dwarfs known to be contaminated by planetary debris, the material accreted by WD J0914+1914 is extremely depleted in the major rock-forming elements magnesium, silicon, calcium and iron with respect to the bulk Earth, and the circumstellar disk at WD J0914+1914 is much larger than

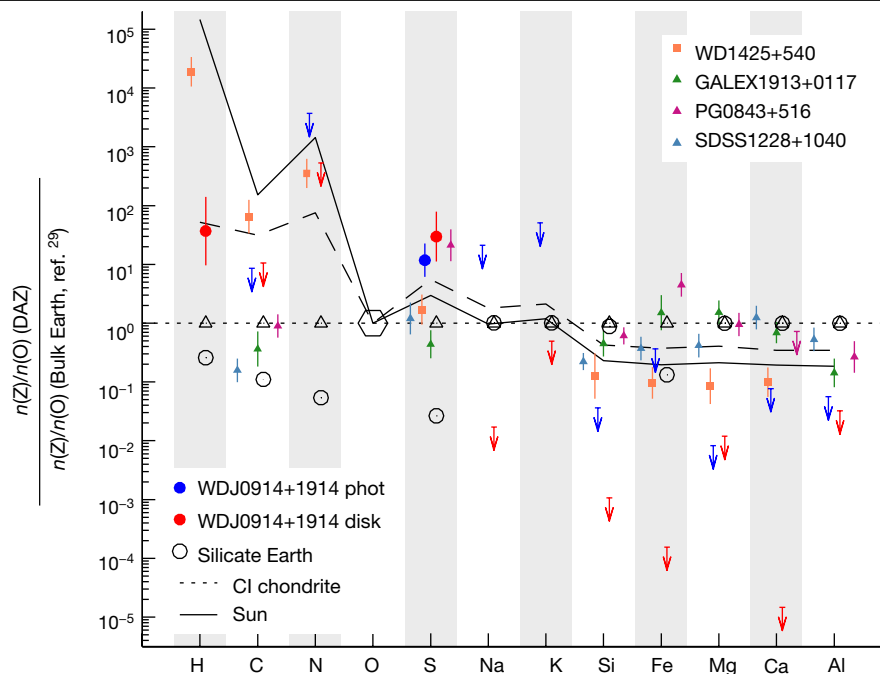
the canonical Roche radius for a rocky body<sup>15</sup>. Both facts argue against tidally disrupted planetesimals<sup>5,16</sup> as the origin of either the gaseous disk, or the photospheric trace metals that we detected. Based on the observational evidence, WD J0914+1914 is a white dwarf accreting from a purely gaseous circumstellar disk, and the most plausible origin of the material in that disk is an evaporating giant planet on a close-in orbit around the white dwarf.

The abundances of WD J0914+1914 are reminiscent of the deeper layers of the ice giants in the Solar System. Modelling the radio and microwave spectrum of Uranus required low concentrations of ammonia (NH<sub>3</sub>) and large concentrations of H<sub>2</sub>O (ref. 17). Condensation of ammonia and hydrogen sulfide (H<sub>2</sub>S) into ammonium hydrosulfide (NH<sub>4</sub>SH) is potentially efficient at removing ammonia from the atmosphere. However, for a solar sulfur-to-nitrogen ratio, there is insufficient sulfur to sequester all NH<sub>3</sub> into NH<sub>4</sub>SH. A plausible model for the spectrum of Uranus required H<sub>2</sub>O and H<sub>2</sub>S concentrations increased by factors of a few hundred with respect to their solar values<sup>17</sup>. H<sub>2</sub>S was recently detected in the atmospheres of Uranus<sup>18</sup> and Neptune<sup>19</sup>, confirming that H<sub>2</sub>S ice is a major constituent of the deeper cloud layers of icy giant planets.



**Fig. 2 | Photospheric oxygen and sulfur lines.** The optical spectrum of WD J0914+1914 contains strong photospheric lines of oxygen (a) and sulfur (b), indicating the ongoing accretion from the circumstellar gas disk. A spectral analysis of these lines results in  $\log(S/O) = -0.9$  (by number).





**Fig. 3 | Abundances of the planetary material at WD J0914+1914.** Shown are the number abundances relative to oxygen, normalized to the corresponding ratio for the bulk Earth<sup>29</sup> and sorted by condensation temperature. The error bars represent  $1\sigma$  uncertainties. The only detected elements are hydrogen (in the circumstellar gas), oxygen and sulfur. Blue dots represent the abundances measured from the analysis of the white dwarf photosphere, red dots represent those derived from the Cloudy photo-ionization model for the circumstellar gas, and the respective upper limits are shown by downward arrows. Included

for comparison are the abundances of the Sun (long dashed lines), CI chondrites (short dashed lines), three white dwarfs accreting rocky debris<sup>16</sup> (triangles, which scatter closely around the bulk Earth abundances) and the one white dwarf accreting a Kuiper-belt-like object<sup>7</sup> (squares, broadly resembling solar abundances). The material at WD J0914+1914 is depleted by orders of magnitude in rock- and dust-forming elements (Si, Fe, Mg, Ca) with respect to all known minor planetary bodies and stars.

Intense high-energy (extreme-ultraviolet, EUV) irradiation of Neptune-mass exo-planets results in the photo-evaporation of their atmospheres. Estimated mass loss rates of warm Neptunes with semi-major axes of a few solar radii reach  $10^8$ – $10^{10}$  g s<sup>-1</sup> (for example, GJ 436b<sup>20</sup> and GJ 3470b<sup>21</sup>), comparable to the accretion rate we derive for WD J0914+1914. The high-energy stellar flux required for driving the mass loss rates of the known warm Neptunes is a few per cent of the total host star luminosity, compatible with the high-energy emission of young stars<sup>22</sup>. Photo-evaporation is also the process most likely to cause the mass loss of the giant planet feeding WD J0914+1914. With the accretion disk extending out to around  $10R_{\odot}$ , the planet is probably located at approximately  $15R_{\odot}$  (see Methods). A large fraction of the luminosity of this moderately hot ( $T_{\text{eff}} \approx 27,750$  K) white dwarf emerges in the EUV, which results in high-energy irradiation of the planet very similar to those of mass-losing warm Neptunes orbiting main-sequence stars. The atmospheric escape rate driven by the EUV flux of WD J0914+1914 may be as high as about  $5 \times 10^{11}$  g s<sup>-1</sup> (see Extended Data Fig. 5 and Methods), exceeding those of the warm Neptunes GJ 436b and GJ 3470b<sup>20,21</sup>.

A fraction of the material escaping the atmosphere of the planet remains gravitationally bound to the white dwarf, forming the circumstellar disk detected in the double-peaked emission lines. From this reservoir, the material eventually accretes onto the white dwarf, resulting in photospheric oxygen and sulfur contamination. A photo-ionization model for the gaseous disk implies a strong depletion of hydrogen, which is expected to be the dominant species in the planet's atmosphere, within the circumstellar disk. In addition to its large EUV luminosity, the hot white dwarf also emits copious amounts of Ly $\alpha$  photons, substantially exceeding the solar Ly $\alpha$  flux (see Extended Data Fig. 6 and Methods). Consequently, the inflow of hydrogen is inhibited by its large cross-section in Ly $\alpha$ , strongly enhancing the abundances of oxygen and sulfur in the circumstellar disk and in the accreted material.

A potential analogue to the planet at WD J0914+1914 is HAT-P-26b, a Neptune-mass planet<sup>23</sup> orbiting a K-type star with a period of 4.26 days. The transmission spectrum of HAT-P-26b exhibits strong H<sub>2</sub>O absorption bands, with no detection of carbon-based species<sup>24</sup>. The carbon abundance<sup>24</sup> in the atmosphere of HAT-P-26b,  $\log(\text{C/O}) < -2$ , is below our detection threshold ( $\log(\text{C/O}) < -1.55$ ; see Methods). A detection of carbon in the photospheric spectrum of WD J0914+1914 will require either substantially deeper optical spectroscopy than our 200-min-long X-Shooter observations or far-ultraviolet spectroscopy of the strong C III 1,175 Å transition. Modelling the spectrum of HAT-P-26b predicts sulfur-based cloud-forming condensates<sup>24</sup>, but these are not directly detected. Despite the high temperature of WD J0914+1914, its small radius ( $0.015R_{\odot}$ ) implies a luminosity that is lower than that of F-, G- or K-type main-sequence host stars. Hence despite the intense EUV irradiation, a planet orbiting the white dwarf WD J0914+1914 will be cooler than an equivalent planet around a main-sequence star.

Gravitational interactions in multi-planet systems can perturb planets onto orbits with pericentres close to the white dwarf, where tidal effects are likely to lead to circularization of the orbit. Common envelope evolution provides an alternative scenario for bringing a planet into a close orbit around the white dwarf<sup>25</sup>, though it requires finely tuned initial conditions and only works for planets more massive than Jupiter (see Methods). As the white dwarf continues to cool, the mass loss rate will gradually decrease, and become undetectable in about 350 million years (see Extended Data Fig. 8). By then, the giant planet will have lost around 0.002 Jupiter masses (or around 0.04 Neptune masses), that is, a very small fraction of its total mass.

The ubiquitous existence of planets around white dwarfs has been indirectly implied by the frequent signatures of planetesimals scattered onto orbits crossing the Roche radii of white dwarfs, with dynamical preference for sub-Jovian mass planets<sup>6,26</sup>. We have inspected all 7,000 or so white dwarfs<sup>27</sup> with SDSS spectroscopy, brighter than  $g=19$

and hotter than 15,000 K for the presence of O I (7,774 Å and 8,446 Å) emission lines, but did not identify another system that resembles WD J0914+1914. Spectroscopic signatures of giant planets at white dwarfs are therefore rare, but follow-up observations of the approximately 260,000 white dwarfs identified with Gaia<sup>27</sup> have the potential to discover a sufficient number of such systems to enable a comparative study of their atmospheric compositions.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1789-8>.

1. Zuckerman, B. & Becklin, E. E. Excess infrared radiation from a white dwarf—an orbiting brown dwarf? *Nature* **330**, 138–140 (1987).
2. Vanderburg, A. et al. A disintegrating minor planet transiting a white dwarf. *Nature* **526**, 546–549 (2015).
3. Koester, D., Gänsicke, B. T. & Farihi, J. The frequency of planetary debris around young white dwarfs. *Astron. Astrophys.* **566**, A34 (2014).
4. Jura, M. A tidally disrupted asteroid around the white dwarf G29–38. *Astrophys. J.* **584**, L91–L94 (2003).
5. Zuckerman, B., Koester, D., Melis, C., Hansen, B. M. & Jura, M. The chemical composition of an extrasolar minor planet. *Astrophys. J.* **671**, 872–877 (2007).
6. Frewen, S. F. N. & Hansen, B. M. S. Eccentric planets and stellar evolution as a cause of polluted white dwarfs. *Mon. Not. R. Astron. Soc.* **439**, 2442–2458 (2014).
7. Xu, S. et al. The chemical composition of an extrasolar Kuiper Belt Object. *Astrophys. J.* **836**, L7 (2017).
8. Gentile Fusillo, N. P., Gänsicke, B. T. & Greiss, S. A photometric selection of white dwarf candidates in Sloan Digital Sky Survey Data Release 10. *Mon. Not. R. Astron. Soc.* **448**, 2260–2274 (2015).
9. Horne, K. & Marsh, T. R. Emission line formation in accretion discs. *Mon. Not. R. Astron. Soc.* **218**, 761–773 (1986).
10. Gänsicke, B. T., Marsh, T. R., Southworth, J. & Rebassa-Mansergas, A. A gaseous metal disk around a white dwarf. *Science* **314**, 1908–1910 (2006).
11. Melis, C. et al. Gaseous material orbiting the polluted, dusty white dwarf HE 1349–2305. *Astrophys. J. Lett.* **751**, 4 (2012).
12. Bauer, E. B. & Bildsten, L. Polluted white dwarfs: mixing regions and diffusion timescales. *Astrophys. J.* **872**, 96 (2019).
13. Ferland, G. J. et al. The 2017 release Cloudy. *Rev. Mex. Astron. Astrofis.* **53**, 385–438 (2017).
14. Pyrzas, S. et al. Post-common envelope binaries from SDSS. XV. Accurate stellar parameters for a cool 0.4 M<sub>\*</sub> white dwarf and a 0.16 M<sub>\*</sub> M dwarf in a 3 h eclipsing binary. *Mon. Not. R. Astron. Soc.* **419**, 817–826 (2012).
15. Davidsson, B. J. R. Tidal splitting and rotational breakup of solid spheres. *Icarus* **142**, 525–535 (1999).
16. Gänsicke, B. T. et al. The chemical diversity of exo-terrestrial planetary debris around white dwarfs. *Mon. Not. R. Astron. Soc.* **424**, 333–347 (2012).
17. de Pater, I., Romani, P. N. & Atreya, S. K. Uranus deep atmosphere revealed. *Icarus* **82**, 288–313 (1989).
18. Irwin, P. G. J. et al. Detection of hydrogen sulfide above the clouds in Uranus's atmosphere. *Nat. Astron.* **2**, 420–427 (2018).
19. Irwin, P. G. J. et al. Probable detection of hydrogen sulphide (H<sub>2</sub>S) in Neptune's atmosphere. *Icarus* **321**, 550–563 (2019).
20. Ehrenreich, D. et al. A giant comet-like cloud of hydrogen escaping the warm Neptune-mass exoplanet GJ 436b. *Nature* **522**, 459–461 (2015).
21. Bourrier, V. et al. Hubble PanCET: an extended upper atmosphere of neutral hydrogen around the warm Neptune GJ 3470b. *Astron. Astrophys.* **620**, A147 (2018).
22. Tu, L., Johnstone, C. P., Güdel, M. & Lammer, H. The extreme ultraviolet and X-ray sun in time: high-energy evolutionary tracks of a solar-like star. *Astron. Astrophys.* **577**, L3 (2015).
23. Hartman, J. D. et al. HAT-P-26b: a low-density Neptune-mass planet transiting a K star. *Astrophys. J.* **728**, 138 (2011).
24. Wakeford, H. R. et al. HAT-P-26b: a Neptune-mass exoplanet with a well-constrained heavy element abundance. *Science* **356**, 628–631 (2017).
25. Nelemans, G. & Tauris, T. M. Formation of undermassive single white dwarfs and the influence of planets on late stellar evolution. *Astron. Astrophys.* **335**, L85–L88 (1998).
26. Mustill, A. J., Villaver, E., Veras, D., Gänsicke, B. T. & Bonsor, A. Unstable low-mass planetary systems as drivers of white dwarf pollution. *Mon. Not. R. Astron. Soc.* **476**, 3939–3955 (2018).
27. Gentile Fusillo, N. P. et al. A Gaia Data Release 2 catalogue of white dwarfs and a comparison with SDSS. *Mon. Not. R. Astron. Soc.* **482**, 4570–4591 (2019).
28. Manser, C. J. et al. Doppler imaging of the planetary debris disc at the white dwarf SDSS J122859.93+104032.9. *Mon. Not. R. Astron. Soc.* **455**, 4467–4478 (2016).
29. McDonough, W. The composition of the Earth. In *Earthquake Thermodynamics and Phase Transformation in the Earth's Interior* (eds Teisseyre, R. & Majewski, E.) 5–24 (Elsevier Science Academic Press, 2000).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019



## Methods

### Discovery and follow-up observations

Two SDSS spectra of WD J0914+1914 were taken in November 2005<sup>30</sup> and March 2012<sup>31</sup> (see Extended Data Fig. 1 and Extended Data Table 1), revealing the H $\alpha$ , oxygen and sulfur emission lines. No noticeable change in the strength of the emission lines is detected between the two epochs.

We observed WD J0914+1914 on 2019 January 12 and 13 using X-Shooter<sup>32</sup> mounted on UT2 of the Very Large Telescope. X-Shooter is a three-armed spectrograph covering the extreme blue (UVB, 330–560 nm), visual (VIS, 560 nm–1  $\mu$ m) and near-infrared (NIR, 1–2.4  $\mu$ m) simultaneously. We obtained ten spectra with 20-min exposure times each. Given the faintness of the star,  $z = 19.9$ , little signal was expected in the NIR arm, and we therefore used the ‘stare’ mode, that is, avoiding nodding. The data were reduced with the Reflex package adopting the standard settings and optimizing the slit integration limits<sup>33</sup>. Finally a weighted average spectrum was computed from the individual UVB and VIS observations. The signal-to-noise ratio of this average spectrum is about 45 and about 55 at 4,300 Å and 7,000 Å, respectively.

The X-Shooter spectrum contains the same emission lines detected in the SDSS spectra, plus several additional oxygen and sulfur lines (Fig. 1). The emission lines are broad and double-peaked, indicating that they originate in a circumstellar disk<sup>9,34</sup>. Also present in the spectrum are multiple strong photospheric absorption lines of oxygen and sulfur, implying ongoing accretion from the disk (Fig. 2). The detection of the emission lines in the 2019 X-Shooter spectra, and comparison with the 2005 SDSS spectrum places a lower limit of 14 years on the life-time of the disk.

### Stellar parameters of the white dwarf and its progenitor

We measured the atmospheric parameters of WD J0914+1914 by fitting pure-hydrogen model spectra<sup>35</sup> to the two SDSS spectra, which are well flux-calibrated. We used the well established technique of fitting the Stark-broadened Balmer line profiles<sup>36,37</sup>, which are sensitive to both temperature and gravity. The total extinction along the line-of-sight towards WD J0914+1914 is low,  $E(B - V) = 0.0305 \pm 0.0006$  (ref. <sup>38</sup>), and normalizing the Balmer lines before the fit effectively removes the effect of extinction. The parameters from the fits to the two SDSS spectra are consistent with each other within the uncertainties, and we take the variance-weighted average as the best-fit values (Extended Data Table 2). Using the cooling models of refs. <sup>39–42</sup>, we computed from the effective temperature,  $T_{\text{eff}} = 27,743 \pm 310$  K and the surface gravity,  $\log(g) = 7.85 \pm 0.06$ , a white dwarf mass of  $M_{\text{wd}} = (0.56 \pm 0.03) M_{\odot}$  and a cooling age of  $13.3 \pm 0.5$  million years. The quoted uncertainties are only of statistical nature. The magnitude of additional systematic uncertainties can, in principle, be assessed from comparing the results from the spectroscopic fit to a joint analysis of the photometry and parallax of the star<sup>43,44</sup>. However, the large parallax uncertainty of WD J0914+1914 (about 22%) severely limits the precision of the atmospheric parameters derived from such an analysis<sup>27</sup>. The spectrophotometric distance implied by our fit is about 625 pc, consistent with the upper limit on the distance based on the Gaia parallax<sup>45</sup>. As an alternative independent test of our spectroscopic fit, we applied an extinction of  $E(B - V) = 0.0305$  to the model spectrum with the atmospheric parameters given above, and then scaled that reddened model to the SDSS  $r$ -band magnitude. We computed a GALEX near-ultraviolet magnitude of 18.07 from this model, which agrees well with the observed value of  $18.06 \pm 0.03$  (ref. <sup>46</sup>).

There is still quite some uncertainty in the low-mass end of the initial-to-final mass relation. Using two different relations results in progenitor masses of about  $1.0 M_{\odot}$  (ref. <sup>47</sup>) and about  $1.6 M_{\odot}$  (ref. <sup>48</sup>). The larger value is in closer agreement with many of the earlier works on the initial-to-final mass relation<sup>49–52</sup>. The main-sequence lifetimes of stars in this mass

range are about 2–10 billion years, that is, the white dwarf cooling age is negligible compared to the total system age.

### Photospheric abundances

Fixing the atmospheric parameters as derived above,  $T_{\text{eff}} = 27,743$  K and  $\log g = 7.85$ , we computed synthetic spectra<sup>35</sup> for a wide range of abundances of C, N, O, Ne, Na, Mg, Al, Si, P, S, Cl, Ar, K, Ca, Sc, Ti, V, Cr, Mn and Fe, and fitted those models to the average X-Shooter spectrum. The only elements detected in the photosphere are oxygen and sulfur at  $\log(\text{O}/\text{H}) = -3.25 \pm 0.20$  and  $\log(\text{S}/\text{H}) = -4.15 \pm 0.20$  (by number), implying  $\log(\text{S}/\text{O}) = -0.9$ , which is far above the solar value of  $-1.57$ , though still within the range of stars within the solar neighbourhood<sup>53</sup>. For all other elements, we derived upper limits (see Extended Data Table 2).

Radiative levitation is negligible for oxygen and sulfur at the effective temperature of WD J0914+1914 (see figure 2 of ref. <sup>54</sup>), and therefore the large photospheric abundances of these elements imply ongoing accretion. Accounting for the diffusion velocities, the photospheric sulfur and oxygen abundances require accretion rates of  $\dot{M}_{\text{S}} = 5.5 \times 10^8 \text{ g s}^{-1}$  and  $\dot{M}_{\text{O}} = 2.7 \times 10^9 \text{ g s}^{-1}$ , respectively. Several studies argue that the gradient of the mean molecular weight resulting from the accretion of metals into the radiative hydrogen atmospheres of warm white dwarfs drives thermohaline mixing<sup>12,55,56</sup>, which would cause the above rates to be underestimated. The most recent studies<sup>12,56</sup> extend only to  $T_{\text{eff}} \approx 20,000$  K, and we conclude that the combined accretion rate of oxygen and sulfur based on purely diffusive sedimentation provides a lower limit of  $\dot{M} \approx 3.3 \times 10^9 \text{ g s}^{-1}$ . The actual rate may be higher by an order of magnitude.

The H $\alpha$  emission line from the circumstellar disk suggests that hydrogen is also accreted onto the white dwarf. However, given that hydrogen is the dominant element in the atmosphere, we are not able to derive the hydrogen fraction in the accreted material. Consequently, the analysis of the photospheric spectrum does not provide a constraint on the contribution of hydrogen to the total accretion rate from the circumstellar disk.

### Dynamical information on the location of the emitting gas

The double-peaked structure of the emission lines arises from the Keplerian motion ( $v_{\text{K}}$ ) of gas in a disk around the white dwarf, with

$$v_{\text{K}} = \sqrt{\frac{GM_{\text{wd}}}{r}} \quad (1)$$

where  $G$  is the gravitational constant, and  $r$  the distance from the centre of the white dwarf. Hence, the morphology of the emission line profiles provides dynamical information on the location of the emitting gas, with the separation of the double peaks corresponding to emission from the outer edge of the disk, and the maximum velocity detected in the line wings corresponding to emission from the inner edge<sup>9,34</sup>. Inspection of the normalized line profiles shows that the morphologies of the individual lines are distinctly different (see Extended Data Fig. 2). In particular, the double-peak separation of the forbidden [S II] lines is narrower than that of H $\alpha$  and O I 8,446 Å, which implies that the region emitting [S II] extends to larger distances from the white dwarf. To estimate the velocity ranges over which the circumstellar gas contributes to the observed emission lines we measured the separation of the double peaks and the maximum extent of the line wings (full width at zero intensity) of H $\alpha$ , O I 8,446 Å and [S II] 4,068 Å (O I 7,774 Å is a relatively widely spaced triplet, which results in more complex sub-structure of the line profile, and the [O I] 6,300, 6,343 Å lines are affected by residuals from the oxygen night-sky airglow of the Earth's atmosphere). Whereas the separation of the double-peaks shows a wide range of velocities (about 150 km s<sup>-1</sup> for [S II] 4,068 Å, about 260 km s<sup>-1</sup> for H $\alpha$  and about 350 km s<sup>-1</sup> for O I 8,446 Å), all lines have similar maximum velocities, about 630–650 km s<sup>-1</sup>, implying that they share a common inner radius in the disk.

Because the inclination of our line of sight against the accretion disk is unknown, the semi-major axes of the Keplerian orbits associated with the velocities measured from the emission lines span a wide range. Adopting a white dwarf mass of  $M_{\text{wd}} = 0.56 M_{\odot}$ , the correspondence between inclination and semi-major axis is illustrated in the Extended Data Fig. 3. Inclinations  $i < 5^{\circ}$  can be excluded because the orbits of the gas would fall inside the white dwarf. For an inclination of  $90^{\circ}$  (edge-on), the inner and outer radii of the gas disk are about  $1R_{\odot}$  and about  $10R_{\odot}$ , respectively.

## A photo-ionization model for the accretion disk

Given the mixture of ionization species seen in emission (H I, O I, S I, S II), the temperature of the circumstellar gas disk is expected to be in the range of about 5,000–10,000 K. Whereas mass transfer through the disk will result in some viscous dissipation, the accretion rate inferred from the photospheric oxygen and sulfur abundances ( $\dot{M} \approx 3.3 \times 10^9 \text{ g s}^{-1}$ ) cannot provide sufficient heating. This problem has been explored and discussed in detail for the known gaseous debris disks around white dwarfs<sup>57</sup>. Instead, photo-ionization by the intense ultraviolet flux from the white dwarf is extremely efficient at heating the upper layers of the disk<sup>58,59</sup>. We used the photo-ionization code Cloudy<sup>13</sup> to develop a simple model that can provide insight into the geometry and the composition of the circumstellar gas at WD J0914+1914.

Cloudy requires the spectral energy distribution, and luminosity of the ionizing source as inputs, for which we computed a white dwarf model spectrum spanning wavelengths from 10 Å to 3 μm with the parameters in the Extended Data Table 1. We adopted the solar abundances for the base composition of the circumstellar gas as provided in solar\_GASS10.abn<sup>60</sup> within the Cloudy distribution.

The geometry of the irradiation of the disk by the white dwarf can broadly be separated into two regimes, depending on the ratio of the disk height to the radius of the white dwarf. The disk height is given by<sup>61</sup>

$$H = \sqrt{k \frac{T_{\text{gas}} r^3}{\mu G M_{\text{wd}}}} \quad (2)$$

where  $k$  is the Boltzmann constant,  $T_{\text{gas}}$  the temperature of the gas and  $\mu$  the mean molecular weight of the gas. The mean molecular weight depends on the abundances of the gas (primarily on the mass fractions of hydrogen, oxygen, and sulfur) and on the degree of ionization, but is not expected to vary much beyond  $\mu \approx (10\text{--}30)m_p$ , where  $m_p$  is the proton mass. The dominant factor in the above expression is therefore the distance  $r$  from the white dwarf, which implies that the disk flares up  $\propto r^{3/2}$ .

Near the white dwarf,  $r \lesssim 1R_{\odot}$ , the disk height is small compared to the radius of the white dwarf, and the disk is illuminated from above. However, owing to the shallow angle,  $\alpha$ , of the incident radiation, the effective path length through the gas is much larger than the actual disk height,  $H/\sin\alpha$ . For distances larger than about  $1R_{\odot}$ , the height of the disk approaches, and eventually exceeds, the radius of the white dwarf, and the assumption of a gas shell illuminated by a point source becomes appropriate. We approximated the near case by a gas shell with a distance  $r$  from the white dwarf, and a thickness  $dr = H/\sin\alpha$ , and the far case by a gas shell with a distance  $r$  from the white dwarf, and  $dr$  as a free parameter.

We computed an initial set of Cloudy models, exploring the following free parameters:  $r$ , the distance from the centre of the white dwarf,  $dr$ , the extent of the gas layer,  $\rho$  the density of the gas, and H/O, the number abundance of hydrogen relative to oxygen. In these initial models, we fixed  $\log(\text{S/O}) = -0.9$ , as determined from the analysis of the white dwarf photospheric spectrum. No elements apart from hydrogen, oxygen and sulfur were included in the model at this stage. The primary input parameter for Cloudy is the hydrogen number density,  $N_{\text{H}}$ , which we computed for a given model from the gas density  $\rho$ , and the H/O and S/O abundance ratios.

The ultraviolet radiation from the white dwarf photo-ionizes the upper layers of the circumstellar disk, heating it to about 10,000–20,000 K. These layers are optically thin in the continuum, and the cooling of the gas takes place via the emission lines detected in the optical spectrum of WD J0914+1914. Deeper layers are essentially neutral, and the observed emission line spectrum does not provide a constraint on the total column density of this neutral material. Within reasonable limits,  $\rho$  and  $dr$  can be traded off against each other, as both parameters determine the total column density of the gas, and hence the total cross-section for intercepting the ultraviolet photons from the white dwarf.

To assess the quality of the Cloudy models, we computed line flux ratios for all observed emission lines, and compared the values from the synthetic spectrum with those measured from the X-Shooter data:

$$Q = \sum_{i=1}^{N_{\text{lines}}} \sum_{j=i+1}^{N_{\text{lines}}} \frac{F_i^{\text{S}}/F_j^{\text{S}}}{F_i^{\text{O}}/F_j^{\text{O}}} + \frac{F_i^{\text{S}}/F_i^{\text{O}}}{F_j^{\text{S}}/F_j^{\text{O}}} \quad (3)$$

where  $F^{\text{O}}$  and  $F^{\text{S}}$  refer to the observed and synthetic line fluxes, respectively. The above function equally penalizes models in which the line fluxes are either too large, or too low.

From the first exploratory models we found that for a solar O/H ratio, the Balmer lines were always much stronger than observed, independent of the exact choice of  $r$ ,  $dr$ , and  $\rho$ . Depleting  $\log(\text{O/H}) \approx 0.29$  resulted in model line flux ratios that were within the correct order of magnitude. At close separations from the white dwarf, low densities ( $\rho \lesssim 10^{-11} \text{ g cm}^{-3}$ ) are insufficient to cool the gas efficiently, and the resulting line flux ratios are incompatible with the observations. For higher densities, cooling becomes more efficient, and the deeper layers are sufficiently cool to produce substantial emission in the O I lines. However, the synthetic spectra contain a number of strong lines that are not observed (O I 3,946 Å, O II 4,650 Å and S I 4,590 Å), and fail to reproduce the line strengths of the observed forbidden lines ([O I], [S II]). In conclusion, this first sequence of models indicated that hydrogen is strongly depleted in the disk, and that geometries corresponding to very low inclinations ( $i \lesssim 20^{\circ}$ ; see Extended Data Fig. 3) that would result in inner disk radii with  $r \ll 1R_{\odot}$  are incompatible with the observations.

To find the parameter space that best reproduces the observed line flux ratios we proceeded to compute a grid of Cloudy models with a fixed  $dr = 0.3R_{\odot}$ , sampling  $0.1R_{\odot} \lesssim r \lesssim 10R_{\odot}$  (constrained by the widths of the observed lines; see Extended Data Fig. 3),  $10^{-12} \text{ g cm}^{-3} < \rho < 10^{-9} \text{ g cm}^{-3}$  and  $\log(\text{O/H}) = -0.11$  to  $0.89$  and  $\log(\text{S/O}) = -1.77$  to  $0.23$ . The quality of the models in the  $(r, \rho)$  plane (Extended Data Fig. 4) illustrates that the best match to the observed line flux ratios is achieved for a location of the gas at  $1R_{\odot} \lesssim r \lesssim 4R_{\odot}$ , and a density of  $\rho \approx 10^{-11.3} \text{ g cm}^{-3}$ . The synthetic spectra in this parameter range produce line flux ratios that are typically consistent with the observed values within a factor of around 2, and do not result in emission lines that are not detected. Combining the constraints from the Cloudy models with those derived from the profile morphology of the observed emission lines (Extended Data Figs. 2, 3) suggests an inclination of the disk  $i \gtrsim 50^{\circ}$ .

## Abundances of the circumstellar disk

The best Cloudy models are found for  $\log(\text{O/H}) \approx 0.29$  and  $\log(\text{S/O}) \approx -0.5$ , with uncertainties of 0.3 dex. For comparison, we derived  $\log(\text{S/O}) = -0.9$  from the photospheric analysis. Both measurements agree within a factor of around 2.5. This is the first instance where the composition of the accreted material is consistently determined by two independent measurements, that is, from the absorption lines within the white dwarf atmosphere, and from the emission lines of the circumstellar gas reservoir.

The fact that the X-Shooter spectrum contains only emission lines of hydrogen, oxygen and sulfur provides upper limits on the abundances of other elements within the circumstellar gas disk that are typically found in white dwarfs accreting planetary debris. Fixing  $r = 2.5R_{\odot}$  and

$\rho = 10^{-11.2} \text{ g cm}^{-3}$ , we proceeded to add additional elements into the disk model, with their initial abundance set to its solar value. The resulting Cloudy spectra predict strong emission lines for C, N, Na, Mg, Al, Si, K, Ca and Fe. We re-computed Cloudy models, reducing the abundances until the line strengths in the models were consistent with the non-detection in the X-Shooter spectrum. The upper limits on the abundances of these elements within the circumstellar gas disk are reported in Extended Data Table 2. Figure 3 illustrates that these upper limits are much more stringent for Na, Si, Fe and Ca compared to the limits obtained from the white dwarf photosphere analysis.

### Emission line profiles from a Keplerian disk

The Cloudy model only takes into account the integrated line fluxes. In order to explore how well this model can also reproduce the observed emission line profiles we convolved the Cloudy spectrum from the computed grid that resulted in the best quality (see Eq. (3)), corresponding to  $r_{\text{in}} = 1.89 R_{\odot}$ ,  $r_{\text{out}} = 0.3 R_{\odot}$ ,  $\rho = 10^{-11.3} \text{ g cm}^{-3}$ ,  $\log(\text{S/O}) = -0.5$ , and  $\log(\text{H/O}) = -0.29$ , with the line profiles of a Keplerian disk. As, at this stage, we are interested in the shape of the line profiles, we normalized the line fluxes of the Cloudy model to those measured from the X-Shooter spectrum, effectively removing the small remaining differences (about a factor of two, see above) in the absolute line fluxes. We used analytical expressions for the Abel transform<sup>34</sup>, a power-law index of zero for the radial intensity distribution, and allowed the inner and outer radii of the disk to vary in order to match the observed emission line profiles. Adopting an inclination of the gaseous disk against the line of sight of  $i = 60^\circ$ , the line widths and separations of the double peaks of H $\alpha$  and O I 8,446 Å are well matched (Extended Data Fig. 2) by inner disk radii of  $r_{\text{in}} \approx (1.0\text{--}1.3) R_{\odot}$  and outer radii of  $r_{\text{out}} \approx (3.0\text{--}3.3) R_{\odot}$ . The more complex structure of the O I 7,774 Å multiplet is also reasonably well reproduced by the same range of  $r_{\text{in}}$  and  $r_{\text{out}}$ . In contrast, [S II] 4,068 Å requires  $r_{\text{in}} \approx (1.0\text{--}1.3) R_{\odot}$  and  $r_{\text{out}} \approx (8\text{--}10) R_{\odot}$ , and the two forbidden [O I] lines also imply similarly large outer radii, even if their double peaks are not well resolved due to the residuals of the sky background subtraction. The larger outer disk radii implied by the line profiles of the forbidden lines confirm the simple estimates we made above (see Extended Data Fig. 3).

Whereas the synthetic line profiles of an axially symmetric disk reproduce the X-Shooter data relatively well, there is a noticeable difference in the shape of the central depression of O I 8,446 Å, with the observations showing a deeper V-shape compared to the U-shape of the model line profile. Similar differences have been observed in the Balmer lines from accretion disks in white dwarf binaries, and have been interpreted as optical depth effects<sup>62</sup>. We also note that matching the observed width of the H $\alpha$  double peaks requires a small amount of additional intrinsic broadening, which could be the result of Stark broadening within the disk<sup>62</sup>.

We conclude that despite our model for the circumstellar gas disk being relatively simple (based on a constant density both in radius and vertical extent of the disk), the overall agreement in both the emerging fluxes and the profile morphology of the emission lines is remarkably good, resulting in a consistent set of parameters both in terms of the geometric location of the gas, and its composition. The reality will have a more complex geometry as well as density gradients. However, including that complexity in the model by introducing additional free parameters is unlikely to provide deeper physical insight.

### Ruling out a stellar / sub-stellar companion

The initial classification of WD J0914+1914 suggested it to be a cataclysmic variable, that is, a short-period binary containing a white dwarf accreting from a Roche-lobe filling low-mass companion. Whereas the double-peaked morphology of the emission lines confirms the presence of a circumstellar gas disk, cataclysmic variables typically have much stronger Balmer (and often helium) lines<sup>63–66</sup>, and no example of a cataclysmic variable with a white dwarf as hot as about 28,000 K dominating the optical spectrum is known<sup>67,68</sup>.

Another class of systems with similar spectroscopic appearance as WD J0914+1914 are detached short-period post-common envelope binaries (PCEBs), that is, white dwarf binaries with low-mass companions, where H $\alpha$  emission from the companion star is commonly detected<sup>69–71</sup>. In PCEBs containing hot white dwarfs, emission lines of calcium and iron originate from the intense irradiation of the companion<sup>72,73</sup>, which are not observed in WD J0914+1914. The emission lines in PCEBs are narrow and single-peaked, and trace the orbital motion of the companion star, with typical periods of hours and radial velocity amplitudes of several 100 km s<sup>-1</sup> (refs. <sup>74,75</sup>). The double-peaked shape of the emission lines in WD J0914+1914 already rules out an origin from an irradiated low-mass companion. Moreover, their velocity variation is less than about 20 km s<sup>-1</sup>, much lower than observed in any of the known PCEBs<sup>75</sup>.

We measured the radial velocity of the white dwarf using ten of the strongest sulfur absorption lines in the X-Shooter UVB spectra. We fixed the relative wavelengths of these lines to their laboratory values, and their width to 1 Å, roughly matching the spectral resolving power, leaving only the depths of the lines, and the white dwarf radial velocity as free parameters. We find a mean white dwarf velocity of  $-47 \text{ km s}^{-1}$  and an average statistical uncertainty of the individual velocity measurements of about 4.0 km s<sup>-1</sup>. In addition, there is a systematic uncertainty arising from imperfections in centring the star in the slit and the instrument model accounting for flexure. We measured this systematic uncertainty from the interstellar Ca K line to be about 3.7 km s<sup>-1</sup>, and added it in quadrature to the statistical uncertainties, resulting in a total uncertainty of the individual radial velocity measurements of about 5.5 km s<sup>-1</sup>. The radial velocities of WD J0914+1914 are consistent with a constant value, that is, the reduced  $\chi^2$  with respect to the mean is  $\chi^2_{\text{red}} = 0.95$ . We conclude that we do not detect a radial velocity variation of the white dwarf, with an upper limit on its radial velocity amplitude of  $K_{\text{wd}} \leq 3 \text{ km s}^{-1}$ . For the typical periods of PCEBs, about 2 h to 1 day<sup>76</sup>, brown dwarf companions are ruled out. In the period range for the mass donating object suggested by our analysis (see below), about 8–10 days, companions with  $M \geq 30 M_{\text{Jup}}$  are ruled out.

Furthermore, a stellar companion would result in an infrared excess with respect to an isolated white dwarf. The location of WD J0914+1914 has been covered by the UKIRT Hemisphere Survey<sup>77</sup> in the *J*-band. WD J0914+1914 is not detected at the  $J = 19.6$  ( $5\sigma$ ) magnitude limit of the UKIRT Hemisphere Survey. The white dwarf alone has  $J = 19.65$ , computed from the synthetic spectrum. Using absolute *J*-band magnitudes of M-dwarfs and L-type brown dwarfs<sup>78</sup>, and a conservative upper limit on the distance of  $d = 631 \text{ pc}$ <sup>45</sup>, the non-detection of WD J0914+1914 in the UKIRT Hemisphere Survey excludes the presence of a companion earlier than an L5 brown dwarf.

The forbidden oxygen and sulfur lines detected in the spectrum of WD J0914+1914 have not been observed in any accreting or detached white dwarf binary. Accretion from the wind of a low-mass companion does result in photospheric metal contamination in these binaries<sup>79</sup>, however, their abundances derived from spectroscopic analysis are broadly consistent with solar abundances of the accreted material, with strong absorption lines of calcium, iron, magnesium and silicon<sup>14,16,80</sup>.

We conclude from the analysis of the observations that WD J0914+1914 is a white dwarf accreting from a circumstellar gas disk with extremely non-solar abundances, and that the origin of the circumstellar disk is not a stellar or brown-dwarf companion.

### Photo-evaporation versus Roche-lobe overflow

The disk size provides constraints on the location of the planet. We assume that the outer radius of disk is traced by the forbidden [S II] and [O I] lines,  $r_{\text{out}} \approx 10 R_{\odot}$ , and that this radius corresponds to the maximum size of the accretion disk allowed by tidal forces, which is approximately 90% of the white dwarf's Roche lobe, that is,  $r_{\text{out}} \approx 0.9 R_{\text{Lwd}}$ <sup>61</sup>. For a given mass of the planet, assuming a circular orbit, and using standard formula for the Roche-lobe radius<sup>81</sup>, this expression allows



an estimate of semi-major axis of the planet's orbit. For Neptune- to Jupiter-mass giant planets this implies semi-major axes of about  $(14\text{--}16)R_{\odot}$  and orbital periods of 8–10 days. We envisage two scenarios in which WD J0914+1914 could accrete from a planet on a close orbit, that is, either via mass loss driven by the intense EUV luminosity of the white dwarf, or via Roche-lobe overflow. Both alternatives are discussed in detail below.

## Photo-evaporation

EUV radiation is known to drive atmospheric mass loss from giant planets in close orbits around their host stars. This hydrodynamic escape is the result of the ionization of hydrogen. In the absence of efficient cooling mechanisms, no hydrostatic solution exists for the atmosphere of a planet subject to intense irradiation, leading to the formation of a trans-sonic flow<sup>82</sup>. Drag forces in this outflow cause heavier elements to be carried with the escaping hydrogen. The detection of Ly $\alpha$  absorption from atomic hydrogen located outside the Roche lobe of the transiting planet HD 209458b clearly demonstrated the escape of atmospheric material from the planet, and provided the first direct evidence for the evaporation of exo-planets<sup>83</sup>. Subsequent Ly $\alpha$  transits were detected in a number of other systems, including the hot Jupiter HD 189733b<sup>84</sup> and the close-in Neptune-mass planet GJ436b<sup>20,85,86</sup>.

In addition to these Ly $\alpha$  transit observations, heavier elements in the extended atmospheres of transiting planets were detected in ultraviolet and X-ray transit spectroscopy<sup>87–89</sup>, showing that the atmospheric escape must be driven by a hydrodynamic process. Apart from the observational detection in a number of individual systems, hydrodynamic escape is thought to play a crucial role in shaping the properties of the population of close-in exoplanets<sup>82</sup>, resulting in the nearly complete absence of Neptune-mass planets with orbital periods of a few days (the warm Neptune desert) as well as the dearth of low-mass planets with 1.5–2 Earth radii (the evaporation valley).

To test the plausibility of hydrodynamic escape for the planet around WD J0914+1914 we determined the EUV flux of the white dwarf and estimated the corresponding evaporation rates using scaling laws derived from detailed hydrodynamic models<sup>82,90</sup>. The incident EUV flux at the position of the planet was obtained by integrating white dwarf model spectra<sup>35</sup> from 10 Å to 912 Å.

Trace metals in the photosphere of WD J0914+1914, resulting from the accretion of planetary material, may block some of the EUV emission<sup>91</sup>. To evaluate the importance of EUV line blanketing, we computed three white dwarf models, fixing the effective temperature and surface gravity to the values determined from the photospheric analysis (Extended Data Table 1): (1) a pure-hydrogen model, (2) a hydrogen model with oxygen and sulfur at the photospheric abundances (Extended Data Table 2), and (3) a hydrogen model including in addition C, N, Na, Mg, Al, Si, P, Cl, Ar, K, Ca, Ti, V, Mn, Fe using the lower of the two upper limits on their abundances (photospheric or disk, Extended Data Table 2) and solar abundances for those elements without meaningful upper limits. We find very small variations of the EUV flux (less than 10%) between the three models, that is, the amount of metal pollution is insufficient to cause much line blanketing. Below, we use model (2), including photospheric sulfur and oxygen. The EUV flux incident upon the planet is shown as a function of orbital separation in the upper panel of Extended Data Fig. 5.

The EUV luminosity of WD J0914+1914 is comparable to that of T Tauri stars which are assumed to efficiently evaporate the atmospheres of their young giant planets. In particular, the atmospheres of Neptunes at separations below 0.1 astronomical units (roughly the outer border of the warm Neptune desert) are supposed to lose large parts of their atmospheres during these early stages. Analogously, the large EUV luminosity of WD J0914+1914 implies that hydrodynamic escape is unavoidable for any planet with a hydrogen-rich atmosphere and a semi-major axis less than  $200R_{\odot}$ .

For large EUV fluxes, hydrodynamic escape can be in the energy-limited or the recombination-limited regime. Hydrodynamic mass

loss scales proportional to the EUV irradiation in the energy-limited regime, and scales with the square root of the EUV irradiation in the recombination-limited regime.

For a Jupiter-mass planet, the transition between both regimes is usually assumed<sup>90</sup> to occur at  $10,000 \text{ erg cm}^{-2} \text{ s}^{-1}$  but can vary depending on the mass and the radius of the planet across a wide range of EUV fluxes<sup>92</sup>, about  $1,000\text{--}100,000 \text{ erg cm}^{-2} \text{ s}^{-1}$ . Given that we currently do not know the mass and radius of the planet at WD J0914+1914, we assume  $10,000 \text{ erg cm}^{-2} \text{ s}^{-1}$  for the transition. Consequently, the mass loss rates we calculate below should be considered as an order-of-magnitude estimate. For the mass loss rate in the energy-limited regime of irradiated giant planets we used<sup>90</sup>:

$$\dot{M} = \frac{\varepsilon \pi F_{\text{EUV}} R_p^3}{G M_p K(\xi)} \quad (4)$$

where  $R_p$  and  $M_p$  are the radius and the mass of the planet,  $F_{\text{EUV}}$  is the incident EUV flux, and  $\varepsilon$  is the efficiency of using the incident energy, which we set to  $\varepsilon = 0.3$  (refs. <sup>90,92</sup>). At close orbital separations, where the Roche lobe ( $R_L$ ) and planet radius become comparable, mass loss is enhanced. This is accounted for by the correction term  $K(\xi = R_L/R_p)$ , for which we used equation (17) from ref. <sup>93</sup>. The mass loss rate driven by the strong EUV irradiation from WD J0914+1914 is shown in the bottom panel of Extended Data Fig. 5. The  $K$ -term is responsible for the steep increase of  $\dot{M}$  towards the smallest separations. For the estimated location of the planet (around  $(14\text{--}16)R_{\odot}$ , grey-shaded region) we obtain a mass loss rate of around  $5 \times 10^{11} \text{ g s}^{-1}$ , depending only weakly on the planet mass. At that distance from the planet, the outflow velocities that are required to reach the Roche lobe of the planet are far smaller than the velocity required to escape the gravitational potential of the white dwarf, and consequently the evaporated material will fall towards the white dwarf.

Hydrogen is probably the dominant species in the planet's atmosphere, driving the hydrodynamic escape, and is hence also expected to be the most abundant element in the circumstellar disk at WD J0914+1914. However, the weakness of H $\alpha$  in the X-Shooter spectrum, compared to the emission lines of oxygen and sulfur already suggests a substantial depletion of hydrogen in the disk with respect to solar abundances, which we quantitatively confirmed with the Cloudy photo-ionization models ( $\log(\text{O}/\text{H}) \approx 0.29$ , compared to the solar ratio of  $-3.31$ ).

Motivated by these results, we explored the effect of forces other than gravity upon the material escaping the atmosphere of the planet. Being a hot white dwarf, WD J0914+1914 is not only bright in the EUV but also in the far-ultraviolet region of the spectrum, where radiation pressure from Ly $\alpha$  photons can transfer momentum to the evaporated hydrogen. This effect is well studied in the Solar System, where the radiation pressure acting upon neutral hydrogen atoms within the heliosphere is proportional to the total flux in the solar Ly $\alpha$  emission line. The relative importance of radiation pressure is usually expressed as  $\mu$ , the ratio between the force related to radiation pressure and gravity, which is very accurately known for the Sun. The effect of Ly $\alpha$  radiation pressure on the motion of interplanetary neutral hydrogen has been measured<sup>94</sup> and depending on the solar cycle  $\mu$  varies between about 0.8 during the minimum of solar activity and about 1.8 during the maximum<sup>95</sup>. Heavier species than hydrogen are much less affected by radiation pressure.

To establish the importance of radiation pressure on the material accreting onto WD J0914+1914 we compared the Ly $\alpha$  flux of the white dwarf with that of the Sun. For that purpose, we retrieved the high-resolution far-ultraviolet spectra of the Sun obtained with the SORCE SOLSTICE instrument<sup>96</sup>. Despite the fact that the white dwarf spectrum shows Ly $\alpha$  in absorption, whereas it is in emission in the spectrum of the Sun, the flux in the very core of Ly $\alpha$  of WD J0914+1914 is comparable to that of the Sun (Extended Data Fig. 7), a simple consequence of

the high effective temperature of the white dwarf. Moreover, the flux in WD J0914+1914 rapidly increases outside the core of Ly $\alpha$ , whereas it drops in the Sun, and as  $M_{\text{wd}} < M_{\odot}$ ,  $\mu \gg 1$ . We conclude that this provides a natural explanation for the low abundance of hydrogen in the circumstellar disk at WD J0914+1914: the strong radiation pressure from the Ly $\alpha$  photons of WD J0914+1914 efficiently inhibits the inflow of hydrogen. This radiation-pressure-driven hydrogen depletion of the material flowing towards the white dwarf results in an accretion rate onto WD J0914+1914 that is much smaller than the estimated mass loss rate (about  $5 \times 10^{11} \text{ g s}^{-1}$ ; see Extended Data Fig. 5). Given that the mass loss rate is an order-of-magnitude estimate only, we conclude that hydrodynamic escape and subsequent accretion of the heavier elements that are dragged by the escaping hydrogen, provides a consistent explanation of our observations.

### Roche-lobe overflow

An alternative possibility for accretion from a giant planet onto WD J0914+1914 is Roche-lobe overflow which can be substantially increased by tidal heating<sup>97</sup>. However, the scenario of Roche-lobe overflow appears to be extremely unlikely for WD J0914+1914 for several reasons. Crucially, the observed emission lines are best reproduced by a circumstellar disk extending up to about  $10R_{\odot}$ , which implies that the planet must be located at  $>10R_{\odot}$  from WD J0914+1914. Even a Jupiter mass planet would have to be substantially inflated (to about 8 Jupiter radii) to fill its Roche lobe at such a large orbital separation. More generally, using an empirical mass-radius relation for giant planets<sup>98</sup>, we find that Roche-lobe overflow should occur at separations of  $(1-2)R_{\odot}$ , clearly incompatible with the derived disk size. Furthermore, the mass transfer rates expected from a Roche-lobe overflow configuration exceed the value we derived from the photospheric analysis by several orders of magnitude. We conclude that Roche-lobe overflow is incompatible with the observational characteristics of WD J0914+1914.

### Common envelope evolution versus planet–planet scattering

Whereas the observational evidence for a giant planet in a close-in orbit around WD J0914+1914 is compelling, it is clear that a planet with an initial semi-major axis of a few tens of solar radii would not have survived the red giant branch evolution of the white dwarf progenitor. The physical mechanism that migrated the planet from several astronomical units onto its current orbit is open to some speculation.

One possibility is common envelope evolution. At the onset of a common envelope, dynamically unstable mass transfer starts from the giant star onto the secondary object, in our case the planet. The timescale for this unstable mass transfer quickly becomes shorter than the thermal timescale of the planet and, as a result, a common envelope forms around the planet and the core of the giant star, the future white dwarf. This common envelope is expelled at the expense of orbital energy, that is, the planet spirals inward.

Common envelope evolution is known to produce binaries containing white dwarfs with stellar<sup>76</sup> and sub-stellar<sup>74,99</sup> companions and orbital periods in the range of hours to days. In fact, common envelope evolution involving planetary mass objects has been suggested as a possible scenario for the formation of low-mass white dwarfs without a detectable stellar companion<sup>25</sup>. As the mass of the planet,  $M_p$ , is much smaller than the white dwarf mass, the final separation after common envelope evolution can be written as<sup>25</sup>:

$$\alpha_f = \frac{\alpha_{\text{CE}} \lambda}{2} \frac{M_{\text{core}} M_p}{M M_{\text{env}}} R_G \quad (5)$$

where  $R_G$  is the radius of the giant star at the beginning of the inspiral phase,  $\alpha_{\text{CE}}$  is the common envelope efficiency,  $\lambda$  is the binding energy parameter, and  $M$  is the mass of the giant star, which can be separated into the core mass (mass of the future white dwarf,  $M_{\text{core}}$ ) and the envelope mass ( $M_{\text{env}}$ ). The latter is going to be expelled during the process.

During common envelope evolution, the planet will move inside the envelope of the giant star and is likely to be completely evaporated. Whether this happens, and at what separation, depends on the temperature structure of the giant star envelope which can be approximated by<sup>25</sup>

$$T \approx 1.78 \times 10^6 \times (r/R_G)^{-0.85} \text{K} \quad (6)$$

The radius at which evaporation of the planet occurs can then be estimated by equating the local sound speed in the envelope and the escape velocity of the planet<sup>100</sup>.

The above approach has previously been used to estimate the outcome of common envelope evolution involving planetary mass companions using constant values for  $\alpha_{\text{CE}}$  and  $\lambda$  (ref. <sup>25</sup>). Throughout the last two decades, however, new constraints on the common envelope efficiency  $\alpha_{\text{CE}}$  have been obtained and algorithms have been developed that calculate the binding energy parameter  $\lambda$ , which has been found to depend sensitively on the mass and radius of the giant star, in particular if recombination energy stored in the envelope is assumed to contribute to expelling the envelope<sup>101,102</sup>. The contributions from recombination energy are usually parameterized with a second efficiency parameter  $\alpha_{\text{rec}}$ .

We calculated the possible outcome of common envelope evolution involving a planetary mass companion taking into account these recent developments. We used the BSE code<sup>103</sup> to compute the evolution of main sequence stars in the range  $(1-8)M_{\odot}$  and determined the binding energy parameter for all core masses close to the mass of WD J0914+1914, that is, we accepted all masses in the range  $(0.55-0.57)M_{\odot}$ . We then used Eq. (5) to determine the final separation for planet masses ranging from super-Earths to the brown dwarf limit. For the planet to survive, the final separation must be sufficiently large that the planet does not evaporate in the red giant envelope, and does not fill its Roche-radius. The latter was calculated from the planet and white dwarf mass and assuming an empirical mass-radius relation for giant planets<sup>98</sup>.

The results derived from these calculations are illustrated in Extended Data Fig. 7 for two different values of the common envelope parameters. First, we used the strict upper limit for the contributions from orbital energy and recombination energy, that is, we assumed that both energies fully contribute to expelling the envelope ( $\alpha_{\text{CE}} = \alpha_{\text{rec}} = 1.0$ ). These calculations provide a stringent upper limit for the final separation (shown as the dashed line in Extended Data Fig. 7). More realistic are smaller values for both efficiencies, for example, observations of white dwarf binaries with M-dwarf stellar companions can be reproduced if  $\alpha_{\text{CE}} = \alpha_{\text{rec}} = 0.25$  (ref. <sup>104</sup>), for which the predicted final separations fall below the solid black line in Extended Data Fig. 7.

The most important conclusion drawn from inspection of Extended Data Fig. 7 is that planets with masses smaller than around  $1M_{\text{Jup}}$  cannot survive common envelope evolution, whereas planets in the mass range of about  $(1-13)M_{\text{Jup}}$  could end up with orbital separations consistent with the estimated location of the planet around WD J0914+1914 at  $(14-16)R_{\odot}$ . In the latter case the initial planet–star separation must have been about 1.5–5 astronomical units (depending on the planet mass) at the onset of mass transfer from the giant star onto the planet, when the giant star was close to the end of its AGB evolution, the binding energy of envelope was smallest and the contributions from recombination energy were largest. Planet population synthesis models predict the fraction of giant planets to increase with stellar mass in agreement with recent observational studies<sup>105</sup>. Most of the white dwarfs in the Galaxy descend from A/F-type stars, and hence their progenitors are likely to have had rich planetary systems. Given that WD J0914+1914 is unique among about 7,000 white dwarfs with similar cooling ages observed by the SDSS, common envelope evolution can plausibly explain the close-in orbit of the planet at WD J0914+1914, but requires it to be more massive than Jupiter.

An alternative scenario explaining the existence of a giant planet in a close-in orbit around WD J0914+1914 is planet-planet scattering.

# Article

Dynamical studies have shown that closely packed planetary systems which remain stable and ordered on the main sequence can become unpacked when the star evolves into a white dwarf<sup>106</sup>. As a consequence of this unpacking, inward incursions of planets can occur throughout the entire white dwarf cooling track for basically all types of planetary masses, ranging from Earth-like objects to giant planets. These inward incursions of planets on largely eccentric orbits will generate strong tidal forces that can circularize the planetary orbit. Planet-planet scattering therefore represents an alternative explanation for the close planet being evaporated by WD J0914+1914, and also works for planet masses lower than the limit for common envelope evolution (about  $1M_{\text{Jup}}$ ).

The large abundance of sulfur in the circumstellar disk at WD J0914+1914 might indicate a planetary mass closer to those of Neptune and Uranus because the fraction of heavier elements is thought to increase with decreasing planetary mass<sup>107</sup>, which would point towards planet-planet scattering causing the inward migration of the planet at WD J0914+1914. However, given the high mass loss rates expected from hydrodynamic escape, we cannot exclude a more massive planet. We therefore conclude that given the currently available observational constraints, both planet-planet scattering as well as common envelope evolution are plausible explanations for the existence of the planet in close orbit around WD J0914+1914.

Additional constraints on the composition of the accreted material from ultraviolet spectroscopy of WD J0914+1914, as well as including tidal effects into  $N$ -body simulations of the evolution of planetary systems around white dwarfs, have the potential to distinguish between the two scenarios.

## The past and future of the planet around WD J0914+1914

As the evolution of white dwarfs is relatively well understood and primarily consists of thermal heat loss through the non-degenerate envelope, and the consequent contraction of this envelope<sup>108</sup>, we can predict the incident flux, and with it the evaporation rate of the planet at WD J0914+1914, and the resulting accretion rate onto the white dwarf, as a function of time. To that end we computed a small grid of white dwarf model spectra covering effective temperatures ranging from 80,000 K to 10,000 K, for the surface gravities corresponding to  $M_{\text{wd}} = 0.56M_{\odot}$  at each temperature. Integrating the EUV fluxes of these model spectra, we then used Eq. (4) to estimate the mass loss rate as a function of effective temperature and cooling age (see Extended Data Fig. 8). As expected, the mass loss rate decreases with time, particularly once the incident flux on the planet drops below  $10,000 \text{ erg cm}^{-2} \text{ s}^{-1}$ , when mass loss becomes directly proportional to the EUV flux. We estimate that accretion of the evaporating material will become undetectable via photospheric metal contamination<sup>3</sup> once the white dwarf has cooled to about 12,000 K, corresponding to a cooling age of around 350 million years, when the mass loss rate drops below  $10^6 \text{ g s}^{-1}$ .

We estimate the total mass loss due to evaporation of the planetary atmosphere by integrating the mass loss rate over the cooling age of the white dwarf, and assuming that the planet reached its current orbit soon after the formation of the white dwarf. The resulting total mass loss is about 0.002 Jupiter masses, or about 0.04 Neptune masses. Thus, hydrodynamic escape will not change the structure of the giant planet around WD J0914+1914.

## Data availability

The SDSS and X-Shooter spectra analysed in this paper are available from the SDSS (<https://www.sdss.org/>) and ESO (<http://archive.eso.org>) archives.

## Code availability

Cloudy is publicly available (<https://www.nublado.org/>). The model atmosphere code of D. Koester is subject to restricted availability.

30. Abazajian, K. N. et al. The Seventh Data Release of the Sloan Digital Sky Survey. *Astrophys. J. Suppl.* **182**, 543–558 (2009).
31. Abolfathi, B. et al. The Fourteenth Data Release of the Sloan Digital Sky Survey: first spectroscopic data from the Extended Baryon Oscillation Spectroscopic Survey and from the Second Phase of the Apache Point Observatory Galactic Evolution Experiment. *Astrophys. J. Suppl.* **235**, 42 (2018).
32. Vernet, J. et al. X-shooter, the new wide band intermediate resolution spectrograph at the ESO Very Large Telescope. *Astron. Astrophys.* **536**, A105 (2011).
33. Freudling, W. et al. Automated data reduction workflows for astronomy. The ESO Reflex environment. *Astron. Astrophys.* **559**, A96 (2013).
34. Smak, J. On the emission lines from rotating gaseous disks. *Acta Astron.* **31**, 395–408 (1981).
35. Koester, D. White dwarf spectra and atmosphere models. *Mem. Soc. Astron. Ital.* **81**, 921–931 (2010).
36. Bergeron, P., Saffer, R. A. & Liebert, J. A spectroscopic determination of the mass distribution of DA white dwarfs. *Astrophys. J.* **394**, 228–247 (1992).
37. Homeier, D. et al. An analysis of DA white dwarfs from the Hamburg quasar survey. *Astron. Astrophys.* **338**, 563–575 (1998).
38. Schlafly, E. F. & Finkbeiner, D. P. Measuring reddening with Sloan Digital Sky Survey stellar spectra and recalibrating SFD. *Astrophys. J.* **737**, 103 (2011).
39. Bergeron, P., Fontaine, G., Tremblay, P.-E. & Kowalski, P. M. Synthetic colors and evolutionary sequences of hydrogen- and helium-atmosphere white dwarfs (2016). <http://www.astro.umontreal.ca/bergeron/CoolingModels/>.
40. Holberg, J. B. & Bergeron, P. Calibration of synthetic photometry using DA white dwarfs. *Astron. J.* **132**, 1221–1233 (2006).
41. Kowalski, P. M. & Saumon, D. Found: the missing blue opacity in atmosphere models of cool hydrogen white dwarfs. *Astrophys. J. Lett.* **651**, 137–140 (2006).
42. Tremblay, P.-E., Bergeron, P. & Gianninas, A. An improved spectroscopic analysis of DA white dwarfs from the Sloan Digital Sky Survey Data Release 4. *Astrophys. J.* **730**, 128 (2011).
43. Tremblay, P.-E. et al. Core crystallization and pile-up in the cooling sequence of evolving white dwarfs. *Nature* **565**, 202–205 (2019).
44. Genest-Beaulieu, C. & Bergeron, P. A comprehensive spectroscopic and photometric analysis of DA and DB white dwarfs from SDSS and Gaia. *Astrophys. J.* **871**, 169 (2019).
45. Bailer-Jones, C. A. L., Rybizki, J., Fouesneau, M., Mantelet, G. & Andrae, R. Estimating distance from parallaxes. IV. Distances to 1.33 billion stars in Gaia Data Release 2. *Astron. J.* **156**, 58 (2018).
46. Bianchi, L. et al. Catalogues of hot white dwarfs in the Milky Way from GALEX's ultraviolet sky surveys: constraining stellar evolution. *Mon. Not. R. Astron. Soc.* **411**, 2770–2791 (2011).
47. Cummings, J. D., Kalirai, J. S., Tremblay, P.-E., Ramirez-Ruiz, E. & Choi, J. The white dwarf initial-final mass relation for progenitor stars from 0.85 to 7.5  $M_{\odot}$ . *Astrophys. J.* **866**, 21 (2018).
48. Kalirai, J. S. et al. The initial-final mass relation: direct constraints at the low-mass end. *Astrophys. J.* **676**, 594–609 (2008).
49. Weidemann, V. Revision of the initial-to-final mass relation. *Astron. Astrophys.* **363**, 647–656 (2000).
50. Catalán, S. et al. The initial-final mass relationship from white dwarfs in common proper motion pairs. *Astron. Astrophys.* **477**, 213–221 (2008).
51. Casewell, S. L. et al. High-resolution optical spectroscopy of Praesepe white dwarfs. *Mon. Not. R. Astron. Soc.* **395**, 1795–1804 (2009).
52. Williams, K. A., Bolte, M. & Koester, D. Probing the lower mass limit for supernova progenitors and the high-mass end of the initial-final mass relation from white dwarfs in the open cluster M35 (NGC 2168). *Astrophys. J.* **693**, 355–369 (2009).
53. Hinkel, N. R., Timmes, F. X., Young, P. A., Pagano, M. D. & Turnbull, M. C. Stellar abundances in the solar neighborhood: the Hyppatia Catalog. *Astron. J.* **148**, 54 (2014).
54. Chayer, P. et al. Improved calculations of the equilibrium abundances of heavy elements supported by radiative levitation in the atmospheres of hot DA white dwarfs. *Astrophys. J.* **454**, 429–441 (1995).
55. Deal, M., Deheuvels, S., Vauclair, G., Vauclair, S. & Wachlin, F. C. Accretion from debris disks onto white dwarfs. Fingering (thermohaline) instability and derived accretion rates. *Astron. Astrophys.* **557**, L12 (2013).
56. Bauer, E. B. & Bildsten, L. Increases to inferred rates of planetesimal accretion due to thermohaline mixing in metal-accreting white dwarfs. *Astrophys. J. Lett.* **859**, 19 (2018).
57. Hartmann, S., Nagel, T., Rauch, T. & Werner, K. Non-LTE models for the gaseous metal component of circumstellar discs around white dwarfs. *Astron. Astrophys.* **530**, A7 (2011).
58. Melis, C., Jura, M., Albert, L., Klein, B. & Zuckerman, B. Echoes of a decaying planetary system: the gaseous and dusty disks surrounding three white dwarfs. *Astrophys. J.* **722**, 1078–1091 (2010).
59. Kinneer, T. *Irradiated Gaseous Discs Around White Dwarfs*. Master's thesis, Univ. of Warwick (2011).
60. Grevesse, N., Asplund, M., Sauval, A. J. & Scott, P. The chemical composition of the Sun. *Astrophys. Space Sci.* **328**, 179–183 (2010).
61. Frank, J., King, A. & Raine, D. J. *Accretion Power in Astrophysics* 3rd edn (Cambridge University Press, 2002).
62. Marsh, T. R. LTE models of the emission lines of the dwarf nova Z Cha. *Mon. Not. R. Astron. Soc.* **228**, 779–796 (1987).
63. Szkody, P. et al. Cataclysmic variables from Sloan Digital Sky Survey. VI. The sixth year (2005). *Astron. J.* **134**, 185–194 (2007).
64. Szkody, P. et al. Finding the instability strip for accreting pulsating white dwarfs from Hubble Space Telescope and optical observations. *Astrophys. J.* **710**, 64–77 (2010).
65. Breedt, E. et al. 1000 cataclysmic variables from the Catalina Real-Time Transient Survey. *Mon. Not. R. Astron. Soc.* **443**, 3174–3207 (2014).
66. Thorstensen, J. R., Alper, E. H. & Weil, K. E. A trip to the cataclysmic binary zoo: detailed follow-up of 35 recently discovered systems. *Astron. J.* **152**, 226 (2016).



67. Gänsicke, B. T. et al. Sdss unveils a population of intrinsically faint cataclysmic variables at the minimum orbital period. *Mon. Not. R. Astron. Soc.* **397**, 2170–2188 (2009).
68. Pala, A. F. et al. Effective temperatures of cataclysmic-variable white dwarfs as a probe of their evolution. *Mon. Not. R. Astron. Soc.* **466**, 2855–2878 (2017).
69. Hillwig, T. C., Honeycutt, R. K. & Robertson, J. W. Post-common-envelope binary stars and the precataclysmic binary PG 1114+187. *Astron. J.* **120**, 1113–1119 (2000).
70. Kawka, A., Vennes, S., Dupuis, J. & Koch, R. The 0.33 day DA plus dMe binary BPM 6502. *Astron. J.* **120**, 3250–3254 (2000).
71. O'Donoghue, D. et al. The DA+dMe eclipsing binary EC13471-1258: its cup runneth over... just. *Mon. Not. R. Astron. Soc.* **345**, 506–528 (2003).
72. Schmidt, G. D., Smith, P. S., Harvey, D. A. & Grauer, A. D. The precataclysmic variable GD 245. *Astron. J.* **110**, 398–404 (1995).
73. Aungwerojwit, A. et al. HS 1857+5144: a hot and young pre-cataclysmic variable. *Astron. Astrophys.* **469**, 297–305 (2007).
74. Maxted, P. F. L., Napiwotzki, R., Dobbie, P. D. & Burleigh, M. R. Survival of a brown dwarf after engulfment by a red giant star. *Nature* **442**, 543–545 (2006).
75. Parsons, S. G. et al. Testing the white dwarf mass-radius relationship with eclipsing binaries. *Mon. Not. R. Astron. Soc.* **470**, 4473–4492 (2017).
76. Nebot Gómez-Morán, A. et al. Post common envelope binaries from SDSS. XII. The orbital period distribution. *Astron. Astrophys.* **536**, A43 (2011).
77. Dye, S. et al. The UKIRT Hemisphere Survey: definition and J-band data release. *Mon. Not. R. Astron. Soc.* **473**, 5113–5125 (2018).
78. Hoard, D. W. et al. Cool companions to white dwarf stars from the Two Micron All Sky Survey All Sky Data Release. *Astron. J.* **134**, 26–42 (2007).
79. Debes, J. H. & Measuring, M. Dwarf winds with DAZ white dwarfs. *Astrophys. J.* **652**, 636–642 (2006).
80. Tappert, C., Gänsicke, B. T., Rebassa-Mansergas, A., Schmidtobreick, L. & Schreiber, M. R. Multiple emission line components in detached post-common-envelope binaries. *Astron. Astrophys.* **531**, A113 (2011).
81. Eggleton, P. P. Approximations to the radii of Roche lobes. *Astrophys. J.* **268**, 368–369 (1983).
82. Owen, J. E. Atmospheric escape and the evolution of close-in exoplanets. *Annu. Rev. Earth Planet. Sci.* **47**, 67–90 (2019).
83. Vidal-Madjar, A. et al. An extended upper atmosphere around the extrasolar planet HD209458b. *Nature* **422**, 143–146 (2003).
84. Lecavelier des Etangs, A. et al. Evaporation of the planet HD 189733b observed in H I Lyman- $\alpha$ . *Astron. Astrophys.* **514**, A72 (2010).
85. Kulow, J. R., France, K., Linsky, J. & Loyd, R. O. P. Ly $\alpha$  transit spectroscopy and the neutral hydrogen tail of the hot Neptune GJ 436b. *Astrophys. J.* **786**, 132 (2014).
86. Lavie, B. et al. The long egress of GJ 436b's giant exosphere. *Astron. Astrophys.* **605**, L7 (2017).
87. Vidal-Madjar, A. et al. Magnesium in the atmosphere of the planet HD 209458 b: observations of the thermosphere-exosphere transition region. *Astron. Astrophys.* **560**, A54 (2013).
88. Ben-Jaffel, L. & Ballester, G. E. Hubble Space Telescope detection of oxygen in the atmosphere of exoplanet HD 189733b. *Astron. Astrophys.* **553**, A52 (2013).
89. Poppenhaeger, K., Schmitt, J. H. M. M. & Wolk, S. J. Transit observations of the hot Jupiter HD 189733b at X-ray wavelengths. *Astrophys. J.* **773**, 62 (2013).
90. Murray-Clay, R. A., Chiang, E. I. & Murray, N. Atmospheric escape from hot Jupiters. *Astrophys. J.* **693**, 23–42 (2009).
91. Chayer, P., Fontaine, G. & Wesemael, F. Radiative levitation in hot white dwarfs: equilibrium theory. *Astrophys. J. Suppl.* **99**, 189–221 (1995).
92. Owen, J. E. & Alvarez, M. A. UV driven evaporation of close-in planets: energy-limited, recombination-limited, and photon-limited flows. *Astrophys. J.* **816**, 34 (2015).
93. Erkaev, N. V. et al. Roche lobe effects on the atmospheric loss from “hot Jupiters”. *Astron. Astrophys.* **472**, 329–334 (2007).
94. Schwadron, N. A. et al. Solar radiation pressure and local interstellar medium flow parameters from Interstellar Boundary Explorer low energy hydrogen measurements. *Astrophys. J.* **775**, 86 (2013).
95. Bzowski, M. et al. Solar parameters for modeling the interplanetary background. In *Cross-Calibration of Far UV Spectra of Solar System Objects and the Heliosphere* (eds Quémerais, E., et al.) 67 (ISSI Scientific Report Series 13, 2013).
96. McClintock, W. E., Rottman, G. J. & Woods, T. N. Solar-Stellar Irradiance Comparison Experiment II (Solstice II): instrument concept and design. *Sol. Phys.* **230**, 225–258 (2005).
97. Valsecchi, F., Rappaport, S., Rasio, F. A., Marchant, P. & Rogers, L. A. Tidally-driven Roche-lobe overflow of hot Jupiters with MESA. *Astrophys. J.* **813**, 101 (2015).
98. Bashi, D., Helled, R., Zucker, S. & Mordasini, C. Two empirical regimes of the planetary mass-radius relation. *Astron. Astrophys.* **604**, A83 (2017).
99. Farihi, J., Parsons, S. G. & Gänsicke, B. T. A circumbinary debris disk in a polluted white dwarf system. *Nat. Astron.* **1**, 0032 (2017).
100. Soker, N. Can planets influence the horizontal branch morphology? *Astron. J.* **116**, 1308–1313 (1998).
101. Dewi, J. D. M. & Tauris, T. M. On the energy equation and efficiency parameter of the common envelope evolution. *Astron. Astrophys.* **360**, 1043–1051 (2000).
102. Zorotovic, M. et al. Post common envelope binaries from SDSS. XIII. Mass dependencies of the orbital period distribution. *Astron. Astrophys.* **536**, L3 (2011).
103. Hurley, J. R., Tout, C. A. & Pols, O. R. Evolution of binary stars and the effect of tides on binary populations. *Mon. Not. R. Astron. Soc.* **329**, 897–928 (2002).
104. Zorotovic, M., Schreiber, M. R., Gänsicke, B. T. & Nebot Gómez-Morán, A. Post-common-envelope binaries from SDSS. IX: constraining the common-envelope efficiency. *Astron. Astrophys.* **520**, A86 (2010).
105. Borgniet, S. et al. Extrasolar planets and brown dwarfs around AF-type stars. X. the SOPHIE sample: combining the SOPHIE and HARPS surveys to compute the close giant planet mass-period distribution around AF-type stars. *Astron. Astrophys.* **621**, A87 (2019).
106. Veras, D. & Gänsicke, B. T. Detectable close-in planets around white dwarfs through late unpacking. *Mon. Not. R. Astron. Soc.* **447**, 1049–1058 (2015).
107. Thorngren, D. & Fortney, J. J. Connecting giant planet atmosphere and interior modeling: constraints on atmospheric metal enrichment. *Astrophys. J. Lett.* **874**, L31 (2019).
108. Fontaine, G., Brassard, P. & Bergeron, P. The potential of white dwarf cosmochronology. *Publ. Astron. Soc. Pacif.* **113**, 409–435 (2001).

**Acknowledgements** Funding for the Sloan Digital Sky Survey IV was provided by the Alfred P. Sloan Foundation, the US Department of Energy Office of Science, and the Participating Institutions. The SDSS website is [www.sdss.org](http://www.sdss.org). Based on observations collected at the European Organisation for Astronomical Research in the Southern Hemisphere under ESO programme 0102.C-0351(A). B.T.G. and C.J.M. were supported by the UK STFC grant ST/P000495. M.R.S. acknowledges support from the Millennium Nucleus for Planet Formation (NPF) and Fondecyt (grant 1181404). O.T. was supported by a Leverhulme Trust Research Project Grant. The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme number 677706 (WD3D).

**Author contributions** All authors contributed to the data interpretation, discussion and writing of this article. B.T.G. wrote the ESO proposal, carried out the observations, and modelled the emission line profiles. M.R.S. developed the models for the past and future evolution of the planet, and for the photo-evaporation. O.T. developed the Cloudy model for the circumstellar disk. O.T. and D.K. carried out the photospheric analysis. N.P.G.F. identified WD J0914+1914 as unusual white dwarf and reduced the X-Shooter data. C.J.M. searched the SDSS spectroscopic data for additional white dwarfs exhibiting oxygen or sulfur lines.

**Competing interests** The authors declare no competing interests.

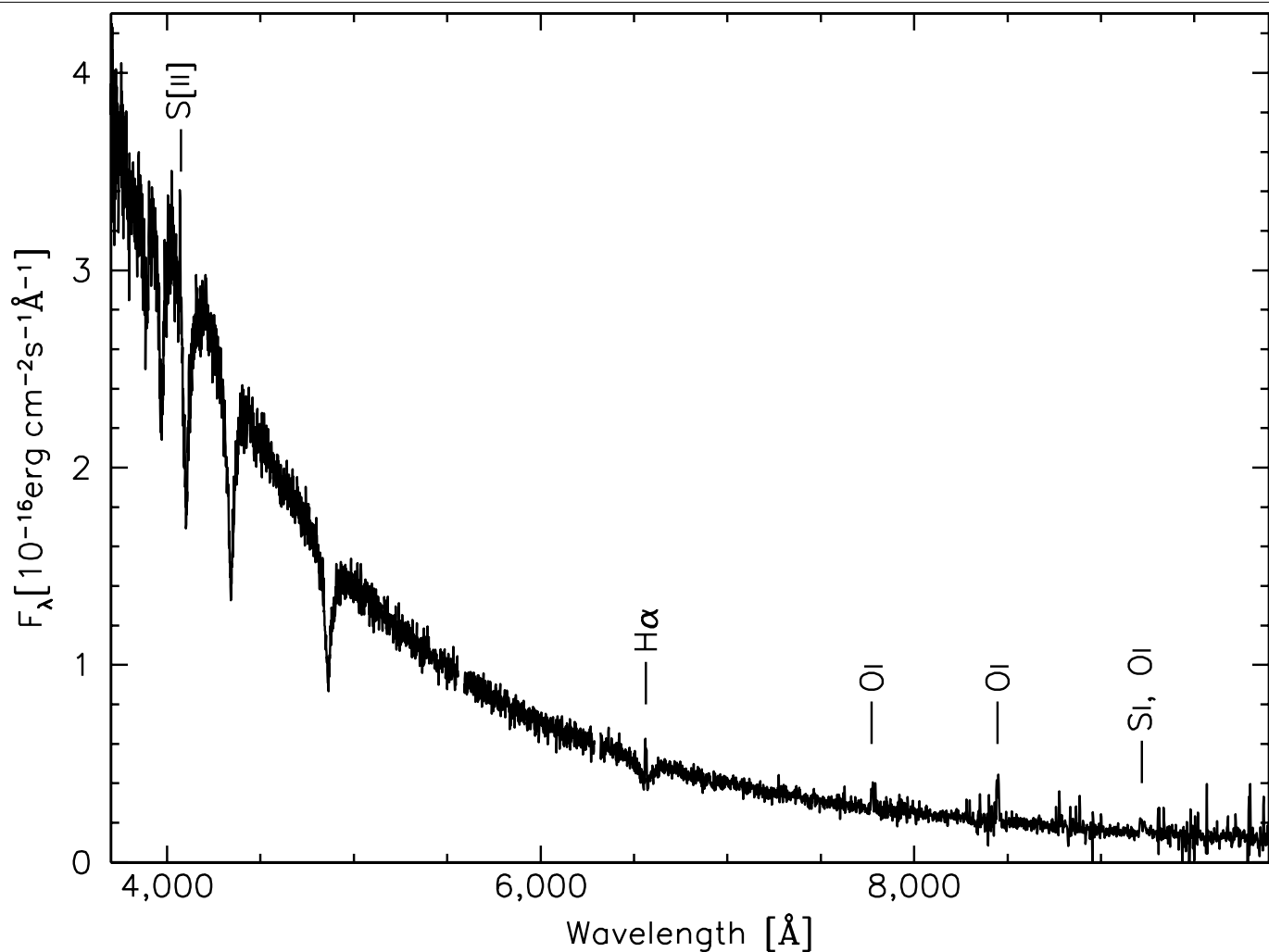
#### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1789-8>.

**Correspondence and requests for materials** should be addressed to B.T.G.

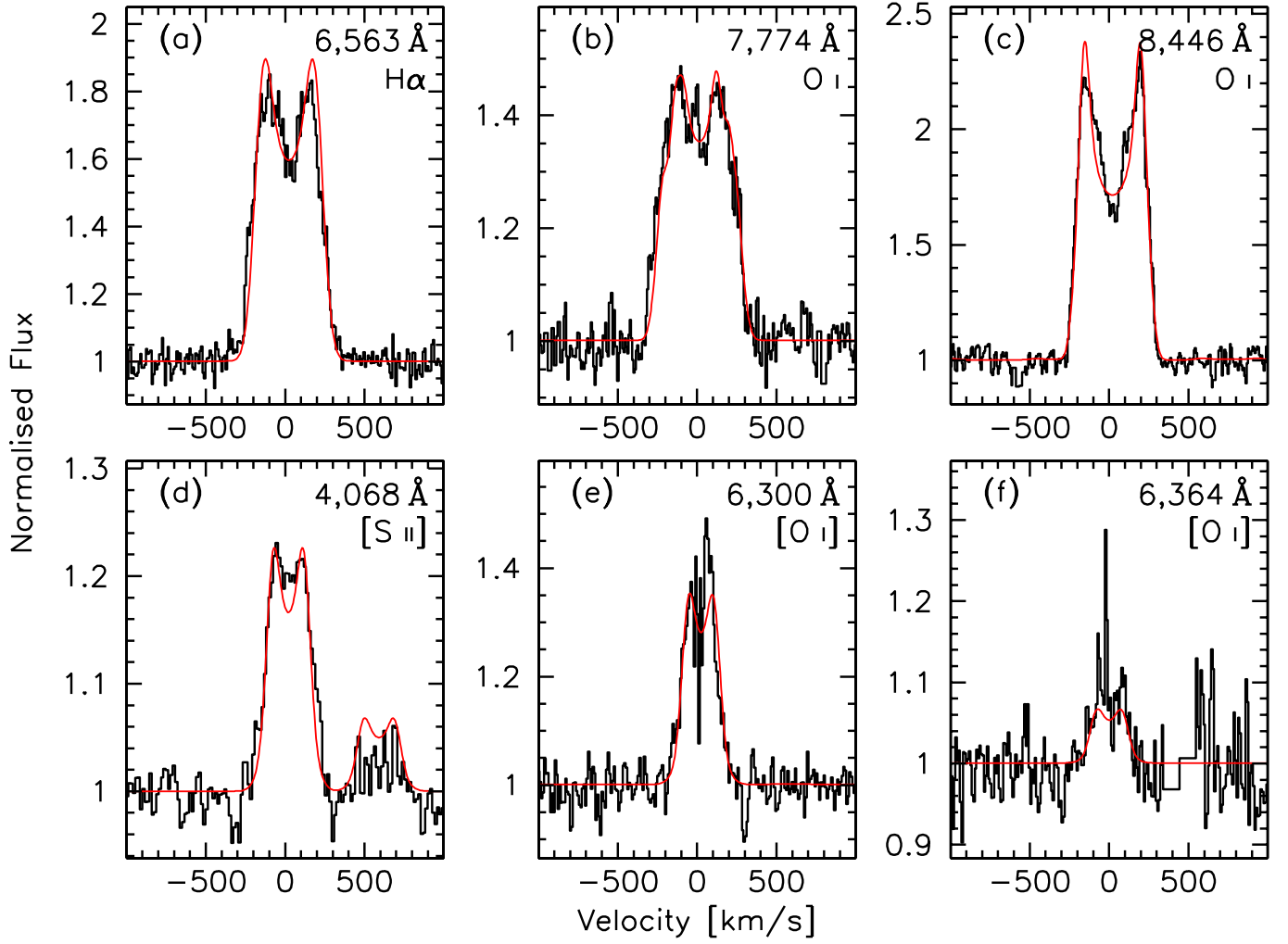
**Peer review information** Nature thanks Patrick Dufour and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



**Extended Data Fig. 1 | Identification spectrum of WD J0914+1914.** The unusual nature of WD J0914+1914 was identified from its optical spectrum within SDSS Data Release 14. The H $\alpha$ , O I 7,774 Å and O I 8,446 Å lines are clearly

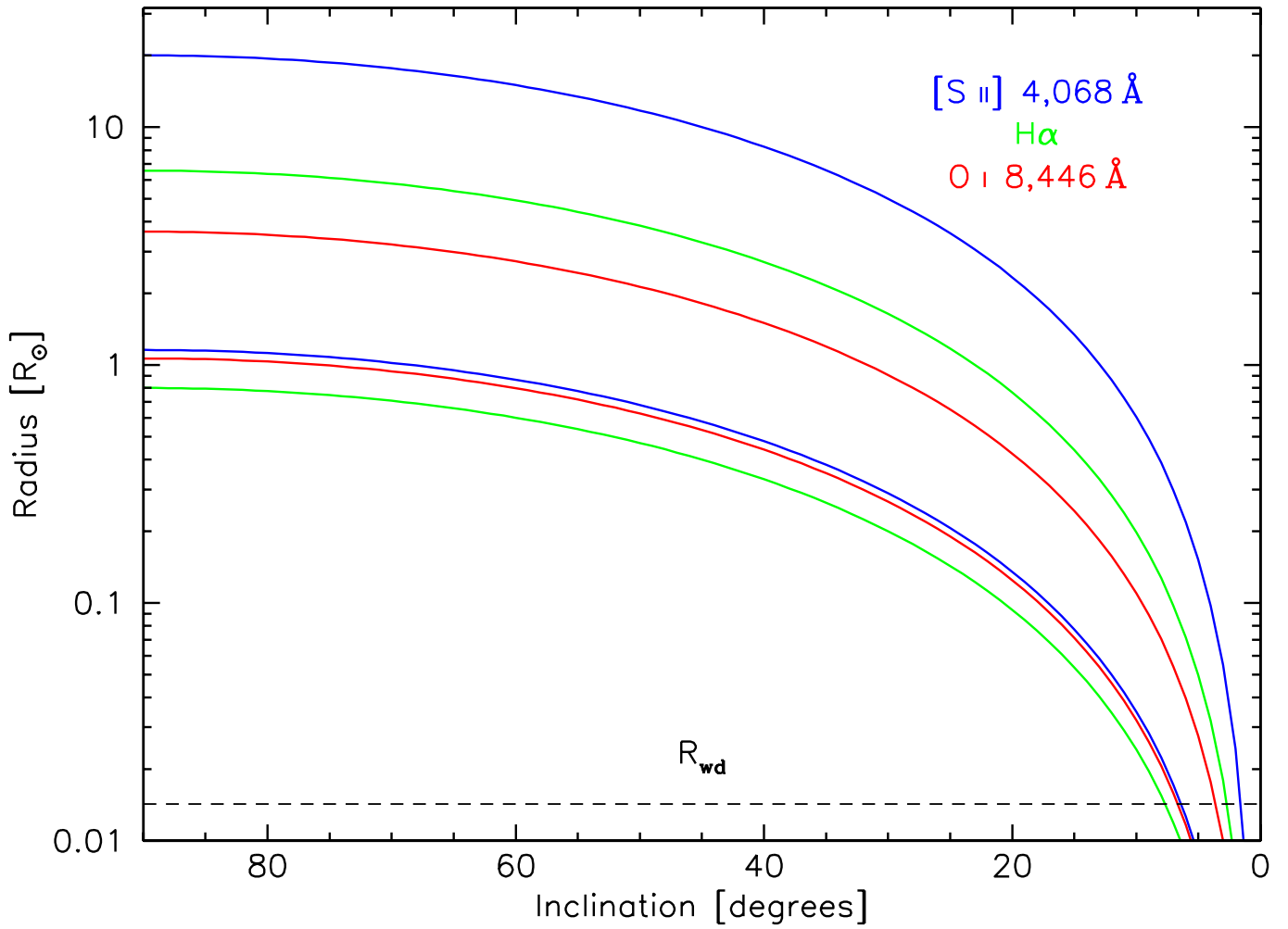
detected, S [II] 4,068 Å and a blend of Si I and O I near 9,240 Å are present near the noise level.



**Extended Data Fig. 2 | Emission lines from a Keplerian disk.** The double-peaked emission lines of hydrogen (a), oxygen (b, c, e, f) and sulfur (d) detected in the optical spectrum of WD J0914+1914 originate in a gaseous circumstellar disk. Shown in red are synthetic disk profiles computed by convolving the Cloudy model that best matches the observed line flux ratios with the broadening function of a Keplerian disk. Adopting an inclination of  $i = 60^\circ$ ,

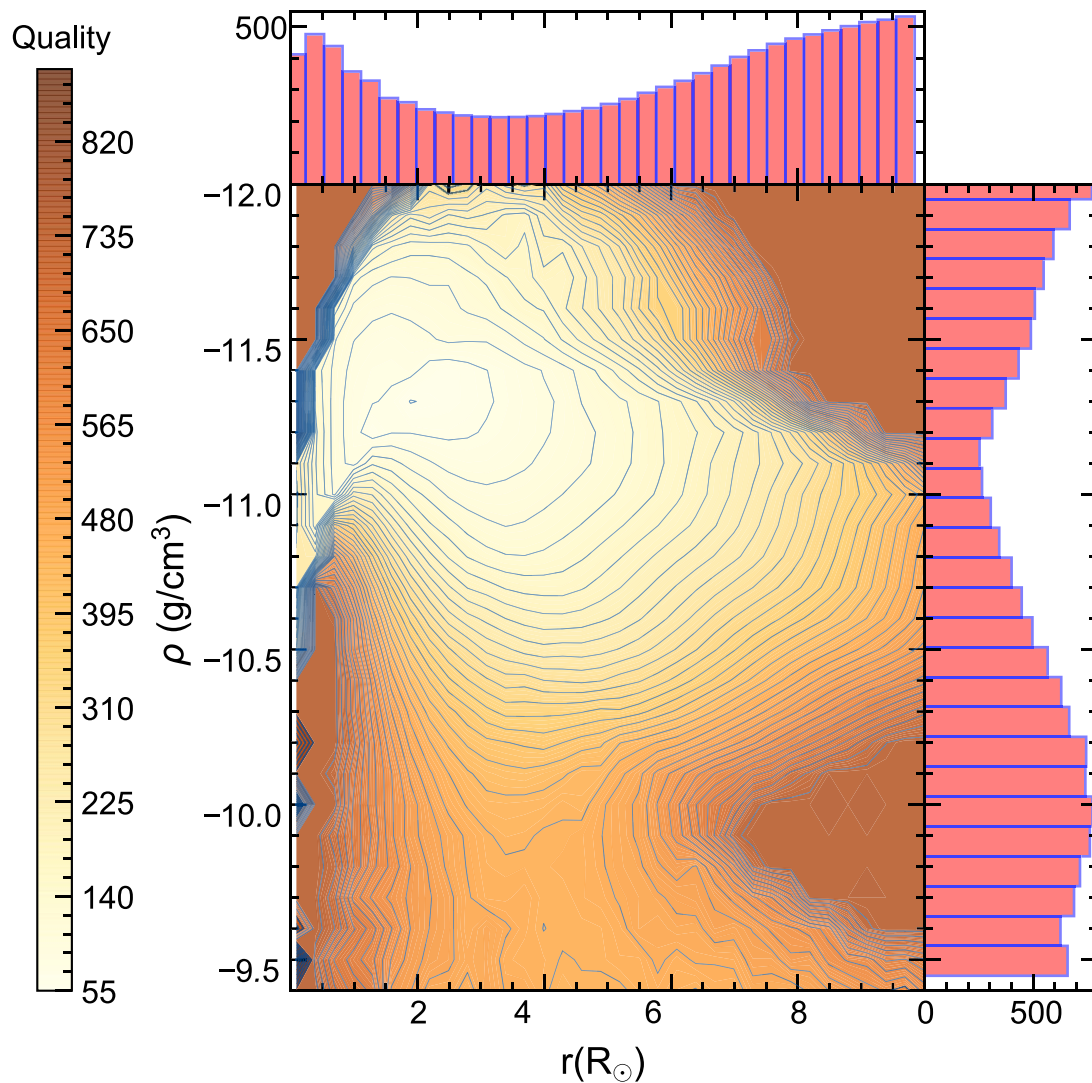
the widths and double-peak separations of the H $\alpha$  (a) and O I 8,446 Å (c) lines are well reproduced for inner and outer disk radii of  $r_{\text{in}} \approx (1.0-1.3)R_\odot$  and  $r_{\text{out}} \approx (2.8-3.3)R_\odot$ , respectively, consistent with the results from the Cloudy models (see Extended Data Fig. 4). The emission of [S II] 4,068 Å (d) extends from about  $1R_\odot$  to  $10R_\odot$ . The V-shaped central depression of the O I 8,446 Å (c) line suggests that the line is optically thick.





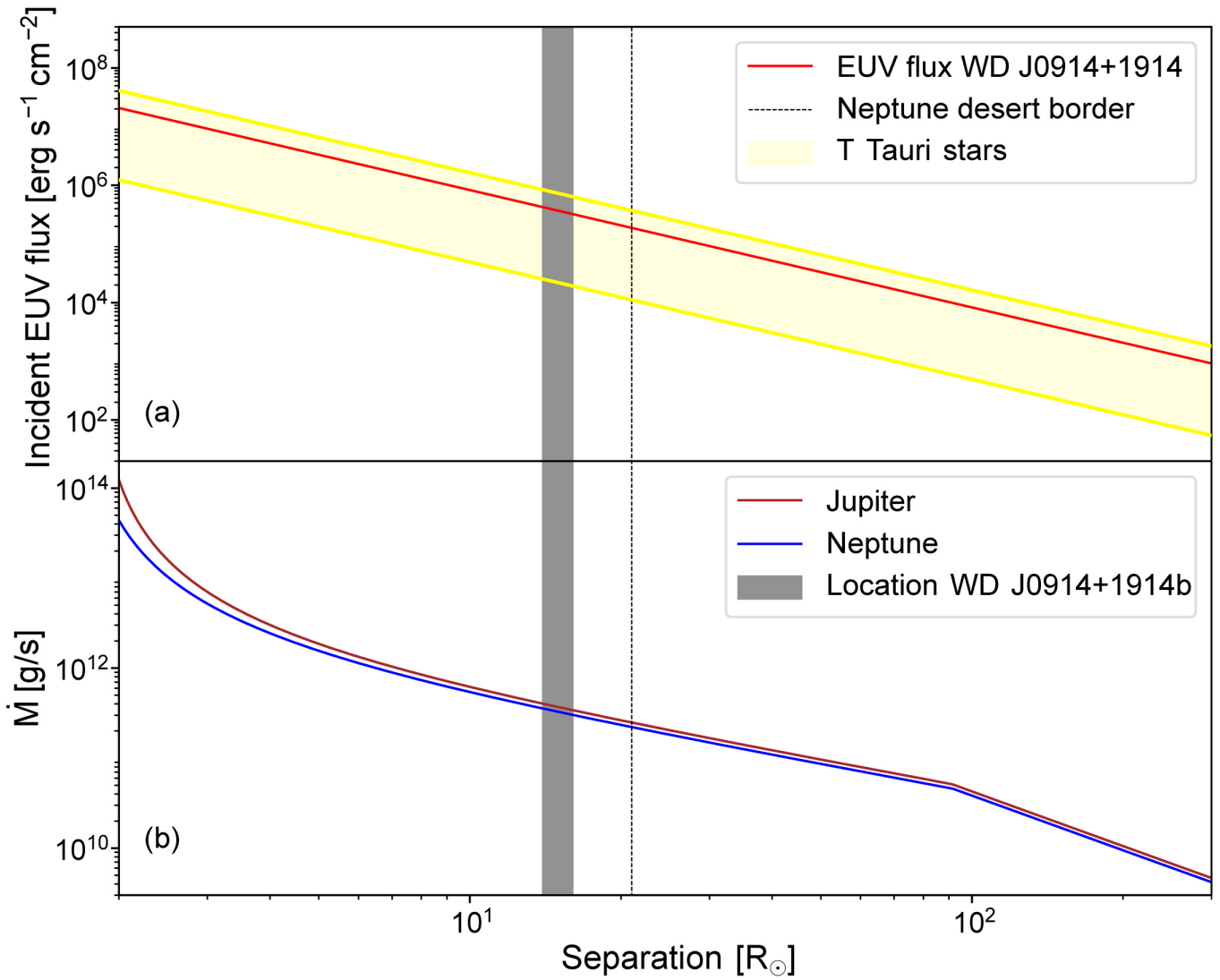
**Extended Data Fig. 3 | Dynamical constraints on the location of the circumstellar gas emitting the observed double-peaked emission lines.** The gas in the circumstellar disk follows Keplerian orbits, and hence the profile shape of the observed emission lines (see Fig. 1 and Extended Data Fig. 2) encodes the location of the gas. The velocity separation of the double-peaks and the maximum velocity in the line wings correspond to motion of gas at the outer edge and inner edge of the disk, respectively. For a given inclination of

the disk, these velocities map into semi-major axes. A lower limit on the inclination,  $i > 5^\circ$ , arises from the finite size of the white dwarf ( $R_{\text{wd}}$ ), and an upper limit on the extent of the disk is provided for an edge-on,  $i = 90^\circ$ , inclination. The forbidden [S II] 4,068 Å line has a much smaller separation of the double-peaks compared to Hα and O I 8,446 Å, implying a larger radial extent.



**Extended Data Fig. 4 | Quality of the Cloudy fits.** The line flux ratios of a grid of Cloudy models spanning a range of gas densities,  $\rho$ , and radial distances from the white dwarf,  $r$ , from the white dwarf are compared to the observed values. The two histograms show the average quality for constant  $r$  (top) and

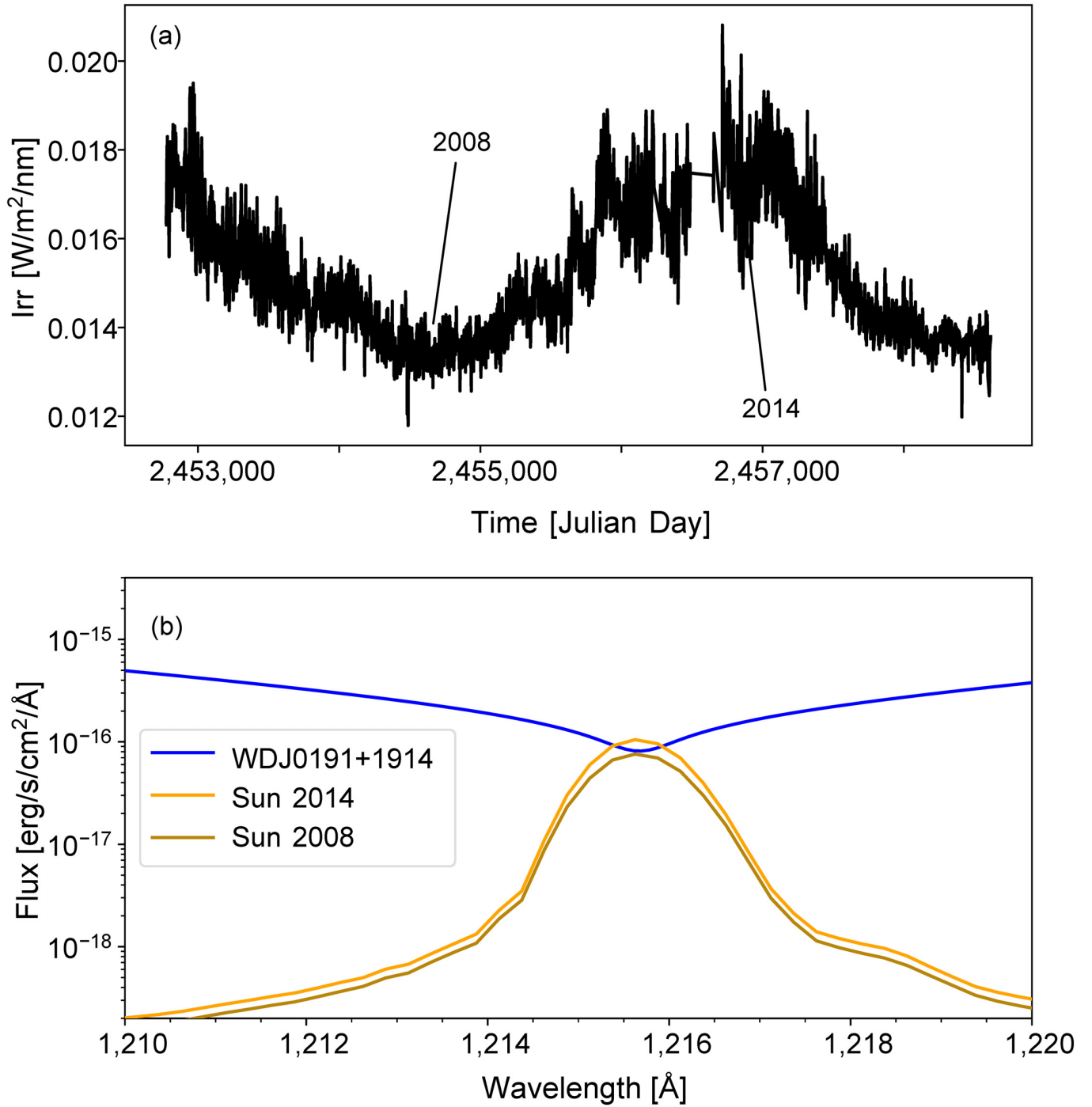
constant  $\rho$  (right). The observed emission line fluxes are reasonably well reproduced by photo-ionized gas with a density of  $\rho = 10^{-11.3} \text{ g cm}^{-3}$  and located at about  $(1-4)R_{\odot}$ .



**Extended Data Fig. 5 | Incident EUV flux and mass loss rates as a function of orbital separation.** **a**, Comparison of the irradiating EUV flux around T Tauri stars (yellow-shaded region) and that of WD J0914+1914 (red line). The outer border of the warm Neptune desert is indicated by the vertical dashed line. The orbital separation of the planet orbiting WD J0914+1914 estimated from the size of the accretion disk is about  $(14\text{--}16)R_{\odot}$  (grey-shaded region). Subject to an EUV luminosity comparable to that of planets around T Tauri stars, the giant

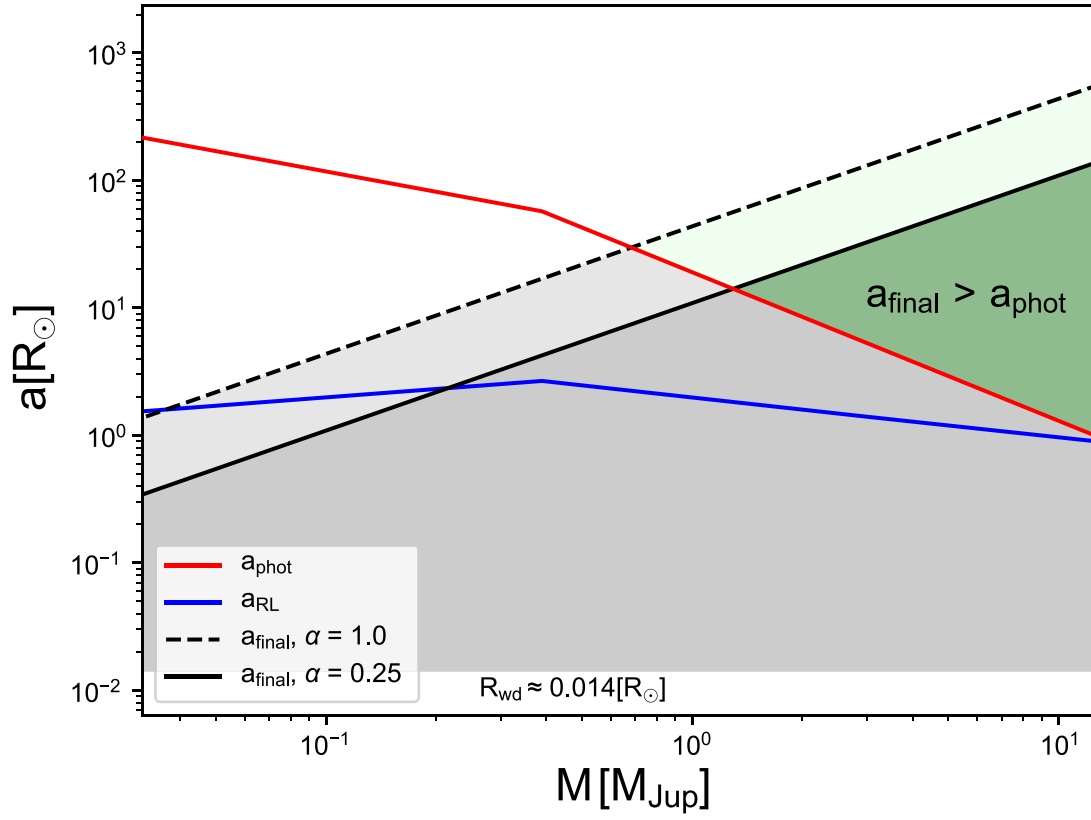
planet at WD J0914+1914 is well within the warm Neptune desert. **b**, Mass loss rates estimated from the assumption of recombination and energy limited hydrodynamic escape for a Jupiter mass and a Neptune mass planet. Substantial mass loss could be generated even for separations of up to a few hundred solar radii, well beyond the estimated orbital location of the giant planet at WD J0914+1914.





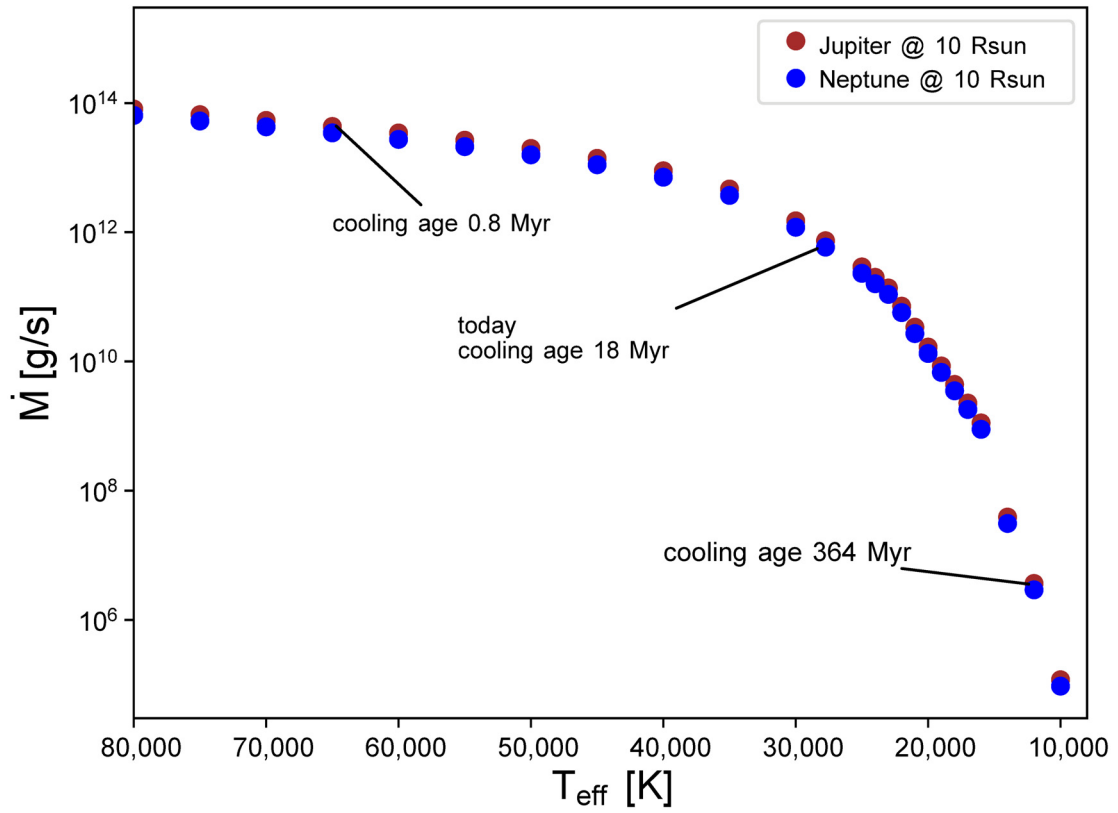
**Extended Data Fig. 6 | Comparison of the the Ly $\alpha$  emission of WD J0914+1914 with the Sun. a,** Ly $\alpha$  irradiance of the Sun across a full solar activity cycle as measured by the SORCE SOLSTICE instrument. The radiation pressure on neutral interplanetary hydrogen in the solar system usually exceeds the gravitational force exerted by the Sun. **b,** The Ly $\alpha$  flux of the Sun during minimum (2008) and maximum (2014) in comparison to the emission of WD

J0914+1914 at a distance of  $15R_{\odot}$ . Given that WD J0914+1914 is less massive than the Sun, and that its Ly $\alpha$  flux is comparable to that of the Sun in the core of the line, but much larger in the wings (even during the 2014 solar maximum), radiation pressure strongly impedes the inflow of hydrogen, explaining the large depletion of hydrogen with respect to oxygen and sulfur within the circumstellar disk.



**Extended Data Fig. 7 | Final separation after common envelope evolution as a function of planetary mass.** We adopted two common envelope efficiencies,  $\alpha = 0.25$  (solid line), and  $\alpha = 1.0$  (dashed line) to calculate an upper limit for the final separation ( $a_{\text{final}}$ ) if the progenitor of WD J0914+1914 and the planet evolved through a common envelope phase. The parameter space of possible outcomes of common envelope evolution lies below these lines (grey-shaded region). We consider the smaller efficiency to be more realistic. For

configurations below the red line ( $a_{\text{phot}}$ ), the planetary mass object will evaporate inside the giant envelope; below the blue line ( $a_{\text{RL}}$ ), it would overflow its Roche lobe. Only planets with parameters within the green-shaded region can survive common envelope evolution. Whereas common envelope evolution can bring a Jupiter-mass planet to the estimated location of the planet around WD J0914+1914 (at  $14\text{--}16 R_{\odot}$ ), smaller planets will be evaporated in the giant envelope.



**Extended Data Fig. 8 | The evolution of the mass loss rate.** White dwarfs cool with time and as a consequence their EUV luminosity decreases. We calculated model spectra for effective temperatures from 80,000 K to 10,000 K, integrated the EUV flux, and determined the mass loss rate of a Jupiter and a Neptune at a distance of  $10R_{\odot}$ . At a cooling age of 364 million years the white

dwarf will have cooled down to 12,000 K, the mass loss rate will drop below about  $10^6 \text{ g s}^{-1}$ , and the resulting photospheric contamination by oxygen and sulfur will become undetectable. Integrating the mass loss rate over the entire cooling time results in a total mass loss of about  $0.002M_{\text{Jup}}$ , which corresponds to about 3.7% of the mass of Neptune.



Extended Data Table 1 | White dwarf parameters

effective temperature $T_{\text{eff}}$ [K]	$27743 \pm 310$
surface gravity $\log g$ [cgs units]	$7.85 \pm 0.06$
white dwarf mass $M_{\text{wd}}$ [ $M_{\odot}$ ]	$0.56 \pm 0.03$
cooling age [Myr]	$13.3 \pm 0.5$
progenitor mass [ $M_{\odot}$ ]	$1.0 - 1.6$
<i>Gaia</i> parallax [milli-arcsec]	$2.17 \pm 0.47$
$u_{\text{SDSS}}$ [mag]	$18.629 \pm 0.026$
$g_{\text{SDSS}}$ [mag]	$18.771 \pm 0.022$
$r_{\text{SDSS}}$ [mag]	$19.198 \pm 0.015$
$i_{\text{SDSS}}$ [mag]	$19.529 \pm 0.022$
$z_{\text{SDSS}}$ [mag]	$19.849 \pm 0.087$
SDSS spectroscopic identifiers	$53700 - 2286 - 0021$
[MJD-PLT-FIBER]	$56017 - 5768 - 0660$

MJD-PLT-FIBER, the modified Julian date and plate and fibre numbers that identify the SDSS spectrum.

Extended Data Table 2 | Element number abundances, log(Z/H)

Element	photosphere	disk	disk scaled	solar
He	< -2.1			-1.07
C	< -4.8	< -1.17	< -4.71	-3.57
N	< -3.7	< -1.00	< -4.54	-4.17
O	-3.25 ± 0.2	0.29 ± 0.3	-3.25	-3.31
Na	< -4.3	< -3.85	< -7.39	-5.76
Mg	< -5.8	< -2.10	< -5.64	-4.40
Al	< -6.0	< -2.70	< -6.24	-5.55
Si	< -5.2	< -3.19	< -6.73	-4.49
S	-4.15 ± 0.2	-0.21 ± 0.3	-3.75	-4.88
K	< -5.2	< -3.67	< -7.21	-6.97
Ca	< -6.0	< -6.18	< -9.72	-5.66
Mn	< -4.4			-6.57
Fe	< -4.2	< -4.03	< -7.57	-4.50
Zn	< -3.0			-7.44

The number abundances in the white dwarf photosphere were derived from fitting model spectra to the oxygen and sulfur lines detected in the X-Shooter spectrum. Upper limits were obtained from the non-detection of the strongest lines of the individual elements. The number abundances in the circumstellar disk were derived from fitting Cloudy models to the observed flux ratios of the emission lines of hydrogen, oxygen and sulfur, and upper limits for the remaining elements were obtained from the non-detection of corresponding emission lines. Hydrogen is strongly depleted in the disk. To facilitate the comparison between these two independent measurements, the column 'disk scaled' gives the abundances and upper limits obtained from the model of the gaseous disk scaled to match the photospheric oxygen abundance. Solar number abundances are provided as reference.

# Observation of the exceptional-point-enhanced Sagnac effect

<https://doi.org/10.1038/s41586-019-1777-z>

Received: 10 April 2019

Accepted: 27 August 2019

Published online: 4 December 2019

Yu-Hung Lai<sup>1,2,6</sup>, Yu-Kun Lu<sup>1,3,4,6</sup>, Myoung-Gyun Suh<sup>1,5,6</sup>, Zhiqian Yuan<sup>1</sup> & Kerry Vahala<sup>1\*</sup>

Exceptional points (EPs) are special spectral degeneracies of non-Hermitian Hamiltonians that govern the dynamics of open systems. At an EP, two or more eigenvalues, and the corresponding eigenstates, coalesce<sup>1–3</sup>. Recently, it was predicted that operation of an optical gyroscope near an EP results in improved response to rotations<sup>4,5</sup>. However, the performance of such a system has not been examined experimentally. Here we introduce a precisely controllable physical system for the study of non-Hermitian physics and nonlinear optics in high-quality-factor microresonators. Because this system dissipatively couples counter-propagating lightwaves within the resonator, it also functions as a sensitive gyroscope for the measurement of rotations. We use our system to investigate the predicted EP-enhanced Sagnac effect<sup>4,5</sup> and observe a four-fold increase in the Sagnac scale factor by directly measuring rotations applied to the resonator. The level of enhancement can be controlled by adjusting the system bias relative to the EP, and modelling results confirm the observed enhancement. Moreover, we characterize the sensitivity of the gyroscope near the EP. Besides verifying EP physics, this work is important for the understanding of optical gyroscopes.

The use of optical microresonators as sensors is being studied across a wide range of applications, including biomolecule<sup>6–8</sup> and nanoparticle<sup>9</sup> detection, temperature measurement<sup>10</sup> and rotation measurement<sup>11–15</sup>. A recently introduced approach to enhancing the response of optical microresonator sensors uses the physics of EPs<sup>4,5,16–20</sup>; operation near an EP boosts the sensor response to a perturbation by an amount that increases with the proximity of the sensor's operating point to the EP<sup>16</sup>. Here we study the EP-induced modification of the Sagnac effect in a microresonator ring laser gyroscope.

The state vectors of a microresonator ring laser gyroscope are admixtures of clockwise (CW) and counter-clockwise (CCW) optical modes and, as will be shown, their location on a Bloch sphere (Fig. 1a) is precisely controllable using Brillouin-induced dispersion<sup>21</sup>. This dispersion is applied independently to the CW and CCW directions using counter-propagating pump waves. Brillouin scattering causes a pump photon with frequency  $\omega_{pj}$  ( $j = 1, 2$ ) to scatter from a co-propagating acoustic phonon with frequency  $\Omega_{\text{phonon}}$  into a backward-propagating Stokes photon. In the resonator, the associated phase-matching condition requires that the Brillouin shift frequency ( $\Omega_{\text{phonon}}$ ) is close in value to an integer multiple ( $\ell$ ) of the free spectral range (FSR) of the resonator, as shown in Fig. 1b (that is,  $\Omega_{\text{phonon}} = \ell \times \text{FSR}$ ). This is achieved by microfabrication control of the resonator diameter and in effect locates a resonator mode (the Stokes mode) within the Brillouin optical-gain spectrum for efficient stimulated Brillouin laser (SBL) action<sup>21,22</sup>. Counter-pumping is performed on the same resonant-mode number ( $m$ ) so that laser action occurs on two counter-propagating Stokes waves belonging to a single mode number (set to  $m - 6$  in this measurement; that is,  $\ell = 6$ ).

To better reveal the non-Hermitian physics of this system, we consider the equation of motion, which reads  $d\psi/dt = H_0\psi$ , where  $\psi = (\alpha_1, \alpha_2)^T$  is the laser mode column vector with amplitudes normalized so that  $|\alpha_{1,2}|^2$  are the photon numbers of the CW and CCW components, respectively, and  $H_0$  is the non-Hermitian Hamiltonian governing the time evolution:

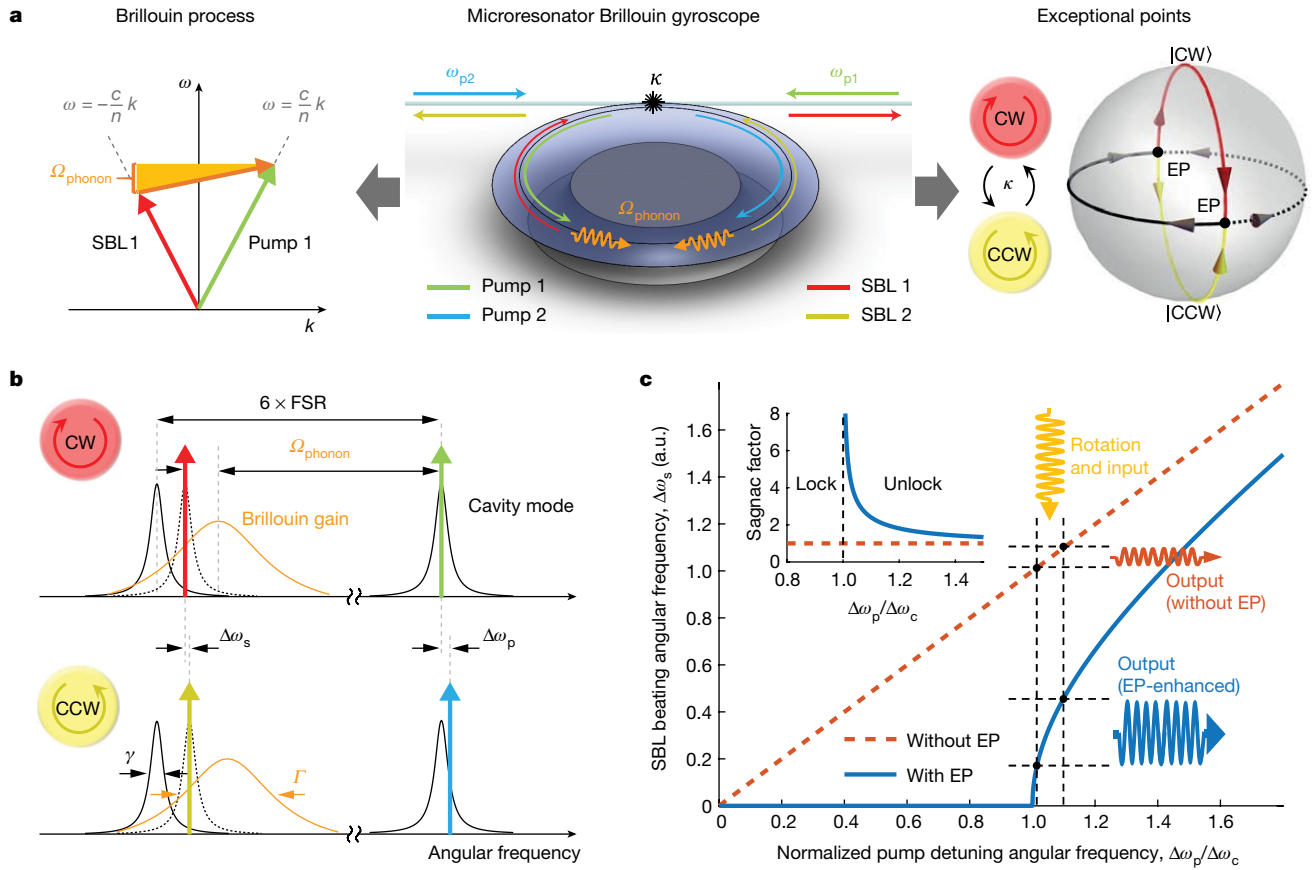
$$H_0 = \begin{pmatrix} \omega_0 + i[g_1 |A_1|^2 - \gamma/2] & i\kappa \\ i\kappa & \omega_0 + i[g_2 |A_2|^2 - \gamma/2] \end{pmatrix} \quad (1)$$

In this expression  $A_1$  and  $A_2$  represent the photon-number-normalized amplitudes of the CCW and CW components of the pump modes, respectively,  $\omega_0$  is the unpumped frequency of the Stokes cavity mode and  $\gamma$  is the cavity damping rate.  $g_j = g_0/[1 + (2i\Delta\Omega_j/\Gamma)]$ , for  $j = 1, 2$ , is the Brillouin gain factor, where  $g_0$  is the gain coefficient,  $\Gamma$  is the gain bandwidth and  $\Delta\Omega_j = \omega_{pj} - \omega_s - \Omega_{\text{phonon}}$  is the frequency mismatch<sup>21</sup>, with  $\omega_s$  the Stokes lasing frequency (an eigenfrequency of equation (1)). The real part of the Brillouin gain factor leads to amplification of the Stokes mode, whereas the imaginary part is responsible for dispersion and consequently mode pulling.  $\kappa$  is the dissipative coupling rate between the two SBL modes and is examined in the Supplementary Information.

In the absence of backscattering ( $\kappa = 0$ ), the CW and CCW SBL processes are independent because the Brillouin gain is intrinsically directional as a result of the phase-matching condition (Fig. 1a). The steady-state lasing condition requires the power loss rate  $\gamma$  to be balanced by the Brillouin gain, which leads to the clamping condition of

<sup>1</sup>T. J. Watson Laboratory of Applied Physics, California Institute of Technology, Pasadena, CA, USA. <sup>2</sup>Present address: OEwaves, Pasadena, CA, USA. <sup>3</sup>Research Laboratory of Electronics, MIT-Harvard Center for Ultracold Atoms, Department of Physics, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>4</sup>Key Laboratory for Mesoscopic Physics and Collaborative Innovation Center of Quantum Matter, School of Physics, Peking University, Beijing, China. <sup>5</sup>Present address: NTT Physics and Informatics Laboratory, East Palo Alto, CA, USA. <sup>6</sup>These authors contributed equally: Yu-Hung Lai, Yu-Kun Lu, Myoung-Gyun Suh. \*e-mail: vahala@caltech.edu





**Fig. 1 | Brillouin control of state vectors in a non-Hermitian system.** **a**, Dual-SBL process in a microresonator. Centre, The green (blue) solid curve represents pump 1 (pump 2) with angular frequency  $\omega_{p1}$  ( $\omega_{p2}$ ) and the red (yellow) solid curve denotes SBL 1 (SBL 2). The orange wavy line represents acoustic phonons with angular frequency  $\Omega_{\text{phonon}}$ . Pumps and output waves are coupled onto the fibre-taper waveguide (light blue). As discussed in Supplementary Information, the backscattering  $\kappa$  is believed to originate primarily from the coupling of the waveguide to the resonator. Left, Brillouin energy and momentum conservation constraints (phase matching) for scattering of a pump wave into a Stokes wave.  $\omega$ , angular frequency of the light;  $k$ , angular wavenumber of the light;  $n$ , refractive index;  $c$ , speed of light in vacuum. Right, CW and CCW modes experience dissipative coupling at rate  $\kappa$ . This coupling creates eigenmodes that are mapped onto a Bloch sphere containing dual EPs (black dots). The trajectories on the Bloch sphere show the evolution of two eigenmodes (red for SBL 1 and yellow for SBL 2) under Brillouin control when the pump detuning frequency decreases from large positive to large negative values. The low-loss and high-loss eigenmodes inside the locking

the pump powers<sup>21</sup>,  $|A_j|^2 = \gamma \frac{1 + (2\Delta\Omega_j/\Gamma)^2}{2g_0}$ . As shown in Supplementary Information, these conditions remain approximately valid for non-zero dissipative backscattering ( $\kappa \neq 0$ ) within the regime in which EP-enhanced rotation measurement is performed (the unlocked regime defined below). As a result, above the lasing threshold equation (1) is simplified to the following form:

$$H_0 = \begin{pmatrix} \omega_0 + \frac{\gamma}{\Gamma}\Delta\Omega_1 & i\kappa \\ i\kappa & \omega_0 + \frac{\gamma}{\Gamma}\Delta\Omega_2 \end{pmatrix} \quad (2)$$

With the introduction of  $\kappa$ , the lasing system exhibits a frequency locking–unlocking transition when varying the pump detuning frequency. The locking regime is known to create a sensing dead band for rotations in ring laser gyroscopes<sup>23</sup>. In the frequency-unlocked regime, the two

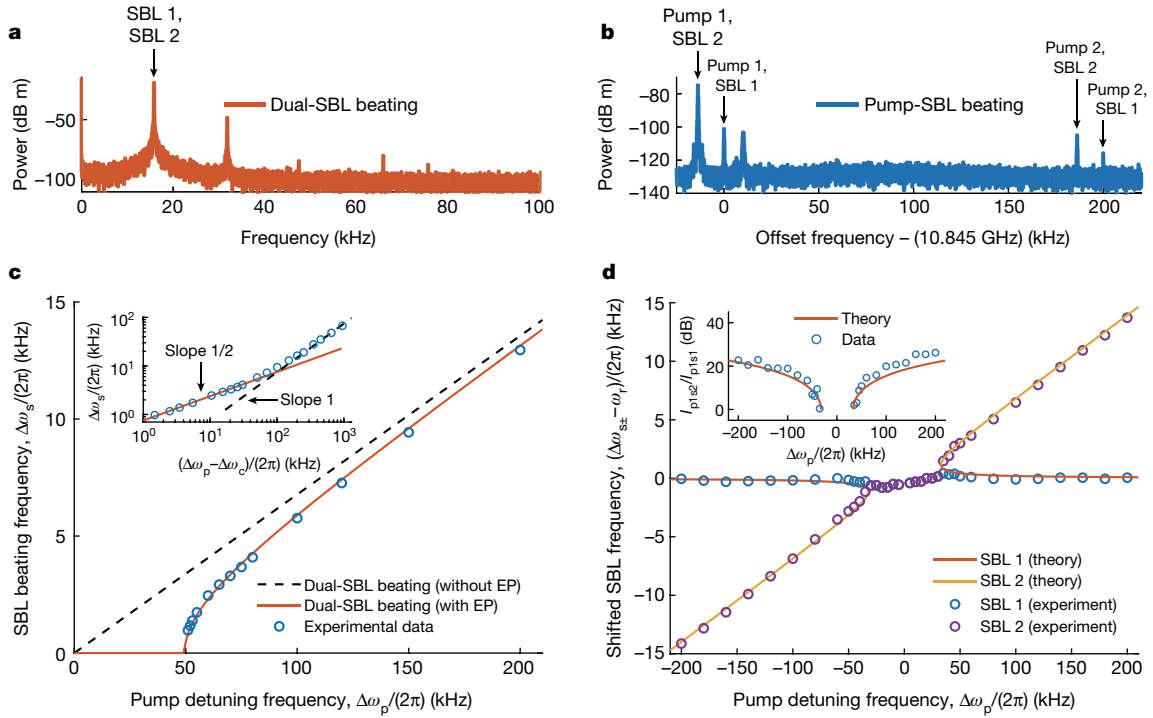
zone are plotted in solid and dashed black curves, respectively (see Supplementary Information for additional discussion). **b**, Efficient laser action requires that each Stokes mode (black; with linewidth  $\gamma$  and separated from the pump by a multiple of the cavity FSR) lies within the Brillouin gain band (orange; with linewidth  $\Gamma$ ), which, through the phase-matching condition, is shifted relative to the pump by  $\Omega_{\text{phonon}} = 4\pi n c_s/\lambda_p$  ( $c_s$ , speed of sound in silica;  $\lambda_p$ , pump wavelength). Here FSR  $\approx 1.8$  GHz, so that  $6 \times \text{FSR}$  approximately matches the Brillouin shift of about 10.8 GHz. Dispersion from the Brillouin gain pulls the Stokes lasing modes by different amounts towards the gain centre on account of the difference in pump angular frequencies,  $\Delta\omega_p$ . **c**, The blue solid curve (red dashed line) shows the dependence of the dual-SBL beating angular frequency,  $\Delta\omega_s$ , on the normalized pump detuning frequency,  $\Delta\omega_p/\Delta\omega_c$  for  $\kappa \neq 0$  ( $\kappa = 0$ ) as per equation (4). The yellow wavy arrow represents the input rotation signal, and the blue (red) solid wavy arrow denotes the output signal with (without) the EP. The inset shows the  $\kappa \neq 0$  Sagnac scale factor normalized to the  $\kappa = 0$  scale factor and shows the enhancement near the EP. a.u., arbitrary units.

lasing modes have distinct angular frequencies  $\omega_{s+}$  and  $\omega_{s-}$ , which are the eigenvalues of the Hamiltonian (equation (2)).

$$\omega_{s\pm} - \omega_r = \frac{\gamma/(2\Gamma)}{1 + \gamma/\Gamma} (\Delta\omega_p \pm \sqrt{\Delta\omega_p^2 - \Delta\omega_c^2}) \quad (3)$$

where  $\omega_r = \{\omega_0 + [\gamma(\omega_{p1} - \Omega_{\text{phonon}})/\Gamma]\}/[1 + \gamma/\Gamma]$ ,  $\Delta\omega_p = \omega_{p2} - \omega_{p1}$  is the pump detuning frequency and  $\Delta\omega_c = 2\Gamma\kappa/\gamma$  is the critical pump frequency detuning at which the system state is at an EP. In deriving this result, it is important to note that the Hamiltonian (equation (2)) depends weakly on its own eigenvalues through  $\Delta\Omega_1$  and  $\Delta\Omega_2$  (see derivation in Supplementary Information). The SBL beating frequency is readily extracted by taking the difference of the above eigenfrequencies,  $\Delta\omega_s = \omega_{s+} - \omega_{s-}$ :

$$\Delta\omega_s = \frac{\gamma/\Gamma}{1 + \gamma/\Gamma} \sqrt{\Delta\omega_p^2 - \Delta\omega_c^2} \quad (4)$$



**Fig. 2 | Measurement of the eigenmode properties.** **a**, Typical measured dual-SBL beating spectrum. **b**, Typical pump-SBL beating spectrum with the frequency axis shifted approximately by 10.845 GHz to centre the pump 1-SBL 1 beating-frequency peak. The individual pump-SBL beating-frequency peaks are identified. **c**, Measured dual-SBL beating frequency versus pump detuning frequency (blue circles). The red solid curve is a fit (with  $\gamma/\Gamma = 0.073$  and  $\kappa = 1.80$  kHz) and the black dotted line corresponds to  $\kappa = 0$  ( $\gamma/\Gamma = 0.073$ ). The data have a slope of 1/2 (slope of 1) near (away from) the EP in the log-log plot in the inset. These data correspond to a mode with larger  $\kappa$  than that used in **d**. **d**, Measured shifted frequencies of the two SBLs,  $(\omega_{s\pm} - \omega_p)/(2\pi)$ , versus the pump

detuning frequency. Theoretical values of  $(\omega_{s+} - \omega_p)/(2\pi)$  and  $(\omega_{s-} - \omega_p)/(2\pi)$  with  $\gamma/\Gamma = 0.076$  and  $\kappa = 1.23$  kHz are plotted as red and yellow lines, respectively. The experimental data of the shifted SBL 1 (SBL 2) frequency are shown as blue (purple) circles. The inset shows the measured power ratio of the CCW components of the lasing modes (blue circles), obtained by the analysis of the spectral components in **b**, and agrees reasonably well with the theoretical prediction (red solid curve).  $I_{\text{plst1}}(I_{\text{plst2}})$  is the strength of the beating signal of pump 1 and SBL 1 (SBL 2). In the main plots of **c** and **d**, the errors in the frequency measurement are typically smaller than the marker size.

This equation is plotted in Fig. 1c. The dissipative coupling between the CW and CCW lasing modes induces second-order EPs at the critical pump-detuning frequencies,  $|\Delta\omega_p| = \Delta\omega_c$ , where the eigenfrequencies and the eigenmodes coalesce. For pump detuning  $|\Delta\omega_p| > \Delta\omega_c$ , the eigenfrequencies bifurcate and the eigenmodes are an unbalanced hybridization of CW and CCW modes. For pump detuning  $|\Delta\omega_p| < \Delta\omega_c$ , the eigenvalues have equal real parts (frequencies) but different imaginary parts (loss rates).

To verify the EP physics predicted above, a high-quality-factor ( $Q \approx 10^8$ ) silica wedge resonator<sup>22</sup> (Fig. 1a) is counter-pumped at distinct frequencies determined by radio-frequency modulation of a single laser (wavelength about 1,552.5 nm). Coupling into the resonator is realized from both ends of a fibre taper<sup>24,25</sup>. One of the pump frequencies is Pound-Drever-Hall-locked to a resonator mode by feedback control on the laser. The second pump frequency is then tuned for state vector control. The two pump powers are stabilized using power feedback. Further details about the experimental setup are provided in Methods.

An electrical spectrum analyser is used to measure the photo-detected dual-SBL beating frequency  $\Delta\omega_s/(2\pi)$  (Fig. 2a) and the SBL-pump beating frequency (Fig. 2b). For the latter, the detection is made along the direction of propagation of pump 1. Plots of these frequencies versus the pump frequency detuning are given in Fig. 2c, d. The provided comparisons with equations (3), (4) show good agreement between theory and measurement. Moreover, the ratio of the CCW components of the eigenmodes is determined from the strength of the beat notes between the CCW pump and the SBL signals (see Supplementary Information for analysis) and is plotted as the inset of Fig. 2d. There is a reasonable agreement between the model and the measurement.

When the resonator experiences an angular rotation rate of  $\Omega$  (positive for the CW direction), the Sagnac effect induces opposing frequency shifts in the CW and CCW modes of a passive resonator such that the differential frequency shift of the CCW mode relative to the CW mode is  $\Delta\omega_{\text{Sagnac}} = 2\pi D\Omega/(n_g\lambda)$ , where  $D$  is the resonator diameter,  $n_g$  is the group index of the passive cavity mode and  $\lambda$  is the laser wavelength<sup>11</sup>. Introducing the opposing frequency shifts (1/2 of the magnitude of the differential shift) as a perturbation into  $H_0$  (equation (2)) modifies the SBL beating frequency as follows:

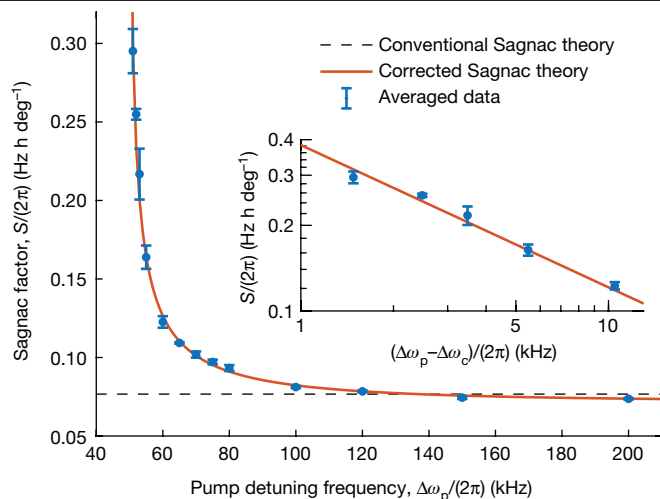
$$\Delta\omega_s = \frac{\gamma/\Gamma}{1 + \gamma/\Gamma} \sqrt{[\Delta\omega_p + \Gamma\Delta\omega_{\text{Sagnac}}/\gamma]^2 - \Delta\omega_c^2} \quad (5)$$

Accordingly, the counter-pumped Brillouin system can function as a gyroscope to measure the rotation signal  $\Omega$  by monitoring the dual-SBL beating frequency  $\Delta\omega_s$ . For comparison with the measurements, the small-signal Sagnac scale factor is now calculated as the derivative of the SBL splitting frequency with respect to the applied rotation rate amplitude  $\Omega$ :

$$S = \frac{\partial\Delta\omega_s}{\partial\Omega}|_{\Omega \rightarrow 0} = \frac{1}{1 + \gamma/\Gamma} \frac{\Delta\omega_p}{\sqrt{\Delta\omega_p^2 - \Delta\omega_c^2}} \frac{2\pi D}{n_g\lambda} \quad (6)$$

where a linear response requires  $\Gamma\Delta\omega_{\text{Sagnac}}/\gamma \ll \Delta\omega_p$ . In this equation, the coefficient  $1/[1 + (\gamma/\Gamma)]$  is a correction resulting from the mode-

pulling effect and  $\Delta\omega_p/\sqrt{\Delta\omega_p^2 - \Delta\omega_c^2}$  is the EP enhancement factor. This enhancement originates from the steep slope of the response curve

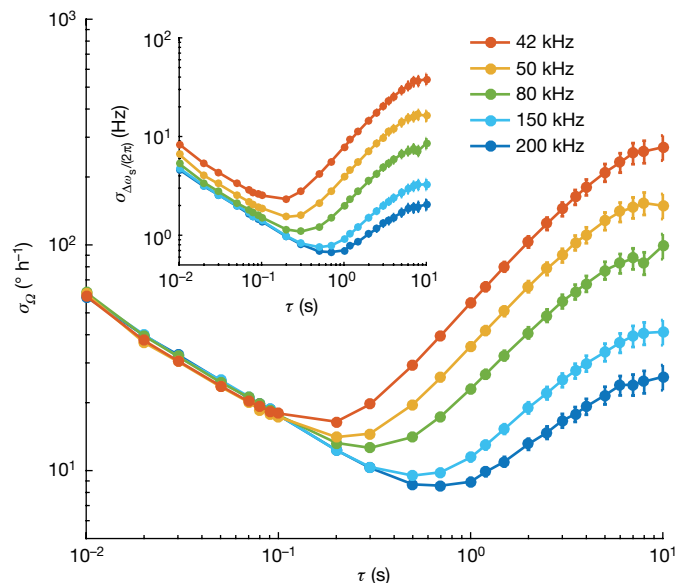


**Fig. 3 | Measured Sagnac scale factor  $S(\Delta\omega_p)$  compared with model results.** The blue dots denote data and the red curve is the theoretical prediction of  $S$  through equation (6). Each point is an average of four measurements and the error bars denote one standard deviation. The Brillouin mode pulling factor,  $1/[1 + (\gamma/\Gamma)]$ , slightly reduces the Sagnac factor at large pump detuning. The black dashed line gives the conventional (without EP enhancement and Brillouin correction) Sagnac factor. The inset shows a log–log plot of five data points near the EP. The red line has a slope of  $-1/2$ , further verifying that the Sagnac enhancement is proportional to  $(\Delta\omega_p^2 - \Delta\omega_c^2)^{-1/2}$ . The approximation  $\Delta\omega_p^2 - \Delta\omega_c^2 \approx 2\Delta\omega_c(\Delta\omega_p - \Delta\omega_c)$  is used for the horizontal axis scale.

near the EP (Fig. 1c) so the scale factor  $S$  surpasses the conventional Sagnac value (that is,  $2\pi D/(n_g\lambda)$ ). Also, we note that the sign of  $S$  depends on the sign of  $\Delta\omega_p$  because the latter determines which SBL wave (CW or CCW) has the higher frequency.

To measure rotations, the resonator is packaged in a small metal box with one edge hinged and the opposing end attached to a piezoelectric stage, in a manner similar to that used in ref. <sup>11</sup>. (As an aside, ref. <sup>11</sup> used a single pump in a cascaded SBL arrangement for rotation sensing. Such an arrangement, however, excludes EP physical effects because the underlying states occur at distinct cavity longitudinal modes.) To create a precise rotation, a sinusoidal oscillation is generated by the piezoelectric stage at a rate of 1 Hz with a fixed amplitude (equivalent to  $410^\circ \text{ h}^{-1}$ ). The resulting time-varying dual-SBL beating frequency is recorded using a frequency counter, and the amplitude of the modulated frequency is extracted by applying a fast Fourier transform to the counter signal. Frequency modulation amplitudes are recorded at several pump frequency detunings. The resulting Sagnac scale factor (that is, the amplitude of the SBL frequency-difference modulation divided by the amplitude of the applied rotation rate) is plotted in Fig. 3. A boost in the Sagnac factor by a factor of 4 compared to the non-EP-enhanced case is observed when operating close to the EP (that is, near the critical detuning frequency). There is good agreement between equation (6) and the measurement, as shown in Fig. 3.

Whereas the Sagnac scale factor is observed to increase near the EP, fluctuation mechanisms exert a greater impact on the measurement, leading to relatively larger errors. To better understand the performance of the gyroscope, we record the beating frequency of the SBLs for up to 600 s without external rotation. The Allan deviation of the beating frequency is then calculated and normalized by the enhanced Sagnac factor  $S(\Delta\omega_p)$  (fitted to the data) so as to arrive at the Allan deviation expressed in rotation-rate units ( $\sigma_a$ ). The results are presented in Fig. 4 for several pump detuning frequencies both near and away from the EP. The smallest pump detuning frequencies are well within the region of EP-enhanced scale factors. At longer times, the Allan deviation data show a drift that is magnified near the EP. This drift is believed to be associated with temperature and power drift in the



**Fig. 4 | Allan deviation of the gyroscope readout at various bias points.** The Allan deviation of the gyroscope readout is measured at several pump detuning frequencies ( $\Delta\omega_p/(2\pi)$ , given in the key) both near (small detuning) and away from (large detuning) the EP. The error bars represent one standard deviation. Inset, Allan deviation of the noise of the SBL beat frequency for the same pump detuning frequencies. In this measurement  $\kappa/(2\pi) = 1.36 \text{ kHz}$  and  $\gamma/\Gamma = 0.0773$ .

system and causes the fluctuations near the EP shown in the data in Fig. 3. At short times, the Allan deviation exhibits angular random walk behaviour and decreases with increasing averaging time. Here, operation in the EP region has little or no impact on the gyroscope sensitivity. To further understand this result, the Allan deviation of the un-normalized frequency noise ( $\sigma_{\Delta\omega_s/(2\pi)}$ ) is plotted in the inset of Fig. 4 and shows that the frequency noise is enhanced in the vicinity of the EP. Indeed, this enhanced noise exactly offsets the scale factor enhancement in the angular random walk regime of the main plot in Fig. 4. The role of technical noise<sup>26</sup> and the consideration of fundamental limits<sup>27–31</sup> to the signal-to-noise ratio of sensors near an EP are recent areas of study, and the source of the enhanced noise in the SBL system is under investigation.

In summary, phase matching of Brillouin gain and dispersion in a microresonator system has been shown to provide precise control of the CW and CCW laser modes near an EP. This control and the inherent high relative stability of the laser modes enable the observation of an enhanced system response to rotations due to the EP-induced modification of the Sagnac effect. By measuring rotations with an approximate amplitude of one revolution per hour, it is possible to observe a four-fold increase of the Sagnac scale factor near the EP. A corresponding sensitivity enhancement with respect to the rotation measurement, as inferred from the measurement of the Allan deviation, was not observed. This work provides a platform for studying EPs in a nonlinear optical system and specifically in the context of rotation sensing.

**Note added in proof.** A study of the enhanced frequency noise in the angular random walk regime in Fig. 4 is presented in ref. <sup>32</sup>.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1777-z>.



1. El-Ganainy, R. et al. Non-Hermitian physics and PT symmetry. *Nat. Phys.* **14**, 11–19 (2018).
2. Feng, L., El-Ganainy, R. & Ge, L. Non-Hermitian photonics based on parity-time symmetry. *Nat. Photon.* **11**, 752–762 (2017).
3. Miri, M.-A. & Alù, A. Exceptional points in optics and photonics. *Science* **363**, eaar7709 (2019).
4. Ren, J. et al. Ultrasensitive micro-scale parity-timesymmetric ring laser gyroscope. *Opt. Lett.* **42**, 1556–1559 (2017).
5. Sunada, S. Large Sagnac frequency splitting in a ring resonator operating at an exceptional point. *Phys. Rev. A* **96**, 033842 (2017).
6. Vollmer, F. & Arnold, S. Whispering-gallery-mode biosensing: label-free detection down to single molecules. *Nat. Methods* **5**, 591–596 (2008).
7. Lu, T. et al. High sensitivity nanoparticle detection using optical microcavities. *Proc. Natl Acad. Sci. USA* **108**, 5976–5979 (2011).
8. Vollmer, F. & Yang, L. Review label-free detection with high-Q microcavities: a review of biosensing mechanisms for integrated devices. *Nanophotonics* **1**, 267–291 (2012).
9. Zhu, J. et al. On-chip single nanoparticle detection and sizing by mode splitting in an ultrahigh-Q microresonator. *Nat. Photon.* **4**, 46–49 (2010); corrigendum **4**, 122 (2010).
10. Xu, X., Jiang, X., Zhao, G. & Yang, L. Phone-sized whispering-gallery microresonator sensing system. *Opt. Express* **24**, 25905–25910 (2016).
11. Li, J., Suh, M.-G. & Vahala, K. J. Microresonator Brillouin gyroscope. *Optica* **4**, 346–348 (2017).
12. Liang, W. et al. Resonant microphotonic gyroscope. *Optica* **4**, 114–117 (2017).
13. Maayani, S. et al. Flying couplers above spinning resonators generate irreversible refraction. *Nature* **558**, 569–572 (2018).
14. Khial, P. P., White, A. D. & Hajimiri, A. Nanophotonic optical gyroscope with reciprocal sensitivity enhancement. *Nat. Photon.* **12**, 671–675 (2018); publisher correction **12**, 714 (2018); author correction **13**, 220 (2019).
15. Gundavarapu, S. et al. Sub-Hertz fundamental linewidth photonic integrated Brillouin laser. *Nat. Photon.* **13**, 60–67 (2019).
16. Wiersig, J. Sensors operating at exceptional points: general theory. *Phys. Rev. A* **93**, 033809 (2016).
17. Wiersig, J. Enhancing the sensitivity of frequency and energy splitting detection by using exceptional points: application to microcavity sensors for single-particle detection. *Phys. Rev. Lett.* **112**, 203901 (2014).
18. Liu, Z.-P. et al. Metrology with PT-symmetric cavities: enhanced sensitivity near the PT-phase transition. *Phys. Rev. Lett.* **117**, 110802 (2016).
19. Hodaei, H. et al. Enhanced sensitivity at higher-order exceptional points. *Nature* **548**, 187–191 (2017); erratum **551**, 658 (2017).
20. Chen, W., Özdemir, Ş. K., Zhao, G., Wiersig, J. & Yang, L. Exceptional points enhance sensing in an optical microcavity. *Nature* **548**, 192–196 (2017).
21. Li, J., Lee, H., Chen, T. & Vahala, K. J. Characterization of a high coherence, Brillouin microcavity laser on silicon. *Opt. Express* **20**, 20170–20180 (2012).
22. Lee, H. et al. Chemically etched ultrahigh-Q wedgeresonator on a silicon chip. *Nat. Photon.* **6**, 369–373 (2012).
23. Chow, W. W. et al. The ring laser gyro. *Rev. Mod. Phys.* **57**, 61–104 (1985).
24. Cai, M., Painter, O. & Vahala, K. J. Observation of critical coupling in a fiber taper to a silica-microsphere whispering-gallery mode system. *Phys. Rev. Lett.* **85**, 74–77 (2000).
25. Spillane, S. M., Kippenberg, T. J., Painter, O. J. & Vahala, K. J. Ideality in a fiber-taper-coupled microresonator system for application to cavity quantum electrodynamics. *Phys. Rev. Lett.* **91**, 043902 (2003).
26. Mortensen, N. A. et al. Fluctuations and noiselimited sensing near the exceptional point of parity-timesymmetric resonator systems. *Optica* **5**, 1342–1346 (2018).
27. Zhang, M. et al. Quantum noise theory of exceptional point amplifying sensors. *Phys. Rev. Lett.* **123**, 180501 (2019).
28. Lau, H.-K. & Clerk, A. A. Fundamental limits and nonreciprocal approaches in non-hermitian quantum sensing. *Nat. Commun.* **9**, 4320 (2018).
29. Langbein, W. No exceptional precision of exceptionalpoint sensors. *Phys. Rev. A* **98**, 023805 (2018).
30. Chen, C., Jin, L. & Liu, R.-B. Sensitivity of parameter estimation near the exceptional point of a non-hermitian system. *New J. Phys.* **21**, 083002 (2019).
31. Pick, A. et al. General theory of spontaneous emission near exceptional points. *Opt. Express* **25**, 12325–12348 (2017).
32. Wang, H., Lai, Y.-H., Yuan, Z., Suh, M.-G. & Vahala, K. J. Petermann-factor limited sensing near an exceptional point. Preprint at <https://arxiv.org/abs/1911.05191> (2019).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

### Experimental setup

In the measurement (see Extended Data Fig. 1), an erbium-doped fibre amplifier is used to boost the power of an external-cavity diode laser, which is split into two arms. Each arm is frequency-shifted by an acousto-optical modulator (AOM) and coupled into the resonator. One arm of the laser is Pound–Drever–Hall-locked to the centre of the cavity mode, and the other arm can be freely tuned by radio-frequency tuning of the AOM. The resonator is shielded passively using insulating foam and the temperature is monitored by a thermistor. Both pump powers are stabilized by active power-feedback control. When changing the pump detuning, the SBL power is almost unchanged (<8%). For the rotation measurement, one corner of the packaged gyroscope is fixed on a pivot point and the other side is placed on a piezoelectric stage so that a precise sinusoidal modulation can be applied.

### Silica wedge resonator

The silica wedge resonator used in this experiment is 36.0 mm in diameter and 8  $\mu\text{m}$  in thickness. The wedge angle is approximately  $30^\circ$ , which is not critical to the measurement. The Brillouin shift in the silica wedge resonator is -10.8 GHz. Details on the fabrication of the silica wedge resonator are provided in ref. <sup>22</sup>.

### Data availability

The data that support the plots within this paper and other findings of this study are available from the corresponding author upon reasonable request.

**Acknowledgements** We thank M. Khajavikhan, D. Christodoulides, O. Peleg and B. Loevsky for discussions during the preparation of this manuscript. We also thank B. Shen, C. Bao and Q. Yang for technical support. Y.-K.L. thanks the Caltech SURF programme for financial support. This project was supported by the Defense Advanced Research Projects Agency (DARPA) under the PRIGM:AIMS programme through SPAWAR (grant number N66001-16-1-4046) and the Kavli Nanoscience Institute.

**Author contributions** Y.-K.L., Y.-H.L., M.-G.S. and K.V. conceived the idea of EP enhancement in the offset-counter-pumped SBL gyroscope. Y.-K.L., Y.-H.L. and K.V. constructed the theoretical model. M.-G.S. fabricated the ultrahigh-Q silica microresonator and helped Y.-H.L. with the packaging. Y.-H.L. and Y.-K.L. performed the experiment. Z.Y. assisted with the gyroscope sensitivity measurements. All authors analysed the data and wrote the manuscript.

**Competing interests** The authors declare no competing interests.

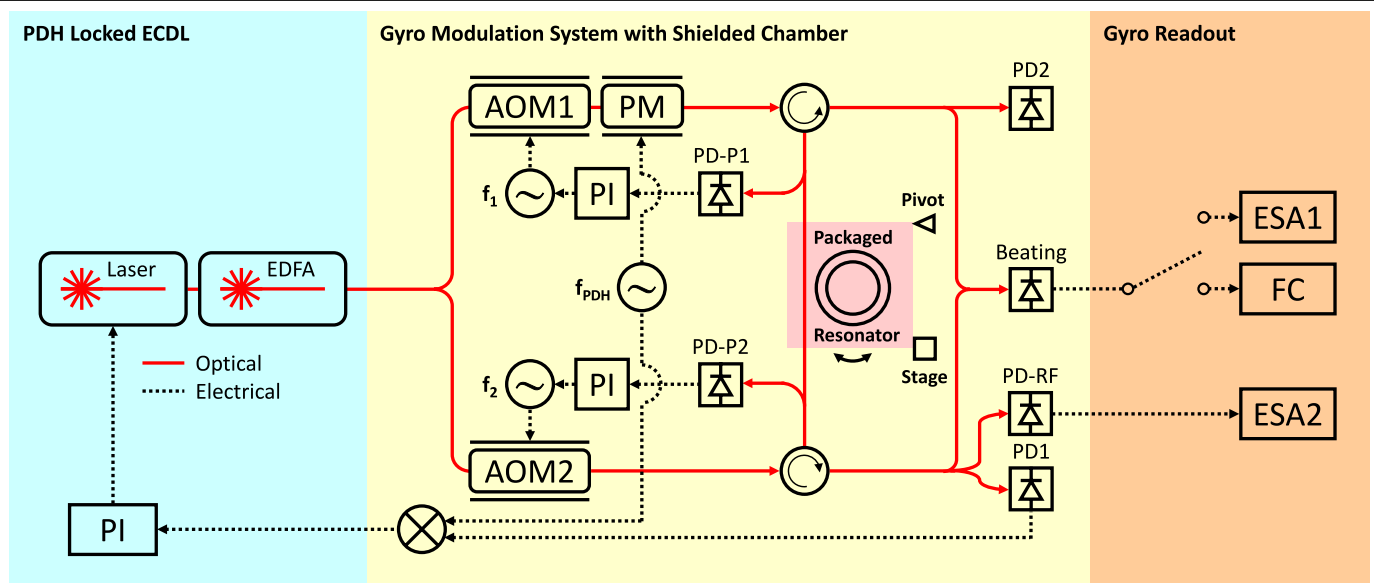
### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1777-z>.

**Correspondence and requests for materials** should be addressed to K.V.

**Peer review information** *Nature* thanks Chia Wei Hsu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



**Extended Data Fig. 1 | Diagram of the counter-pumped SBL gyroscope.**

See Methods for operational description. PDH, Pound–Drever–Hall lock; ECDL, external-cavity diode laser; EDFA, erbium-doped fibre amplifier; PM, phase modulator; PD, photodetector; ESA, electrical spectrum analyser;

FC, frequency counter; PI, proportional-integral servo; RF: radio frequency;  $f_1$  ( $f_2$ ): modulation frequency of AOM1 (AOM2);  $f_{PDH}$ , phase-modulation frequency of the PDH loop.



# Non-Hermitian ring laser gyroscopes with enhanced Sagnac sensitivity

<https://doi.org/10.1038/s41586-019-1780-4>

Received: 10 April 2019

Accepted: 4 September 2019

Published online: 4 December 2019

Mohammad P. Hokmabadi<sup>1</sup>, Alexander Schumer<sup>1,2</sup>, Demetrios N. Christodoulides<sup>1</sup> & Mercedesh Khajavikhan<sup>1,3\*</sup>

Gyroscopes are essential to many diverse applications associated with navigation, positioning and inertial sensing<sup>1</sup>. In general, most optical gyroscopes rely on the Sagnac effect—a relativistically induced phase shift that scales linearly with the rotational velocity<sup>2,3</sup>. In ring laser gyroscopes (RLGs), this shift manifests as a resonance splitting in the emission spectrum, which can be detected as a beat frequency<sup>4</sup>. The need for ever more precise RLGs has fuelled research activities aimed at boosting the sensitivity of RLGs beyond the limits dictated by geometrical constraints, including attempts to use either dispersive or nonlinear effects<sup>5–8</sup>. Here we establish and experimentally demonstrate a method using non-Hermitian singularities, or exceptional points, to enhance the Sagnac scale factor<sup>9–13</sup>. By exploiting the increased rotational sensitivity of RLGs in the vicinity of an exceptional point, we enhance the resonance splitting by up to a factor of 20. Our results pave the way towards the next generation of ultrasensitive and compact RLGs and provide a practical approach for the development of other classes of integrated sensor.

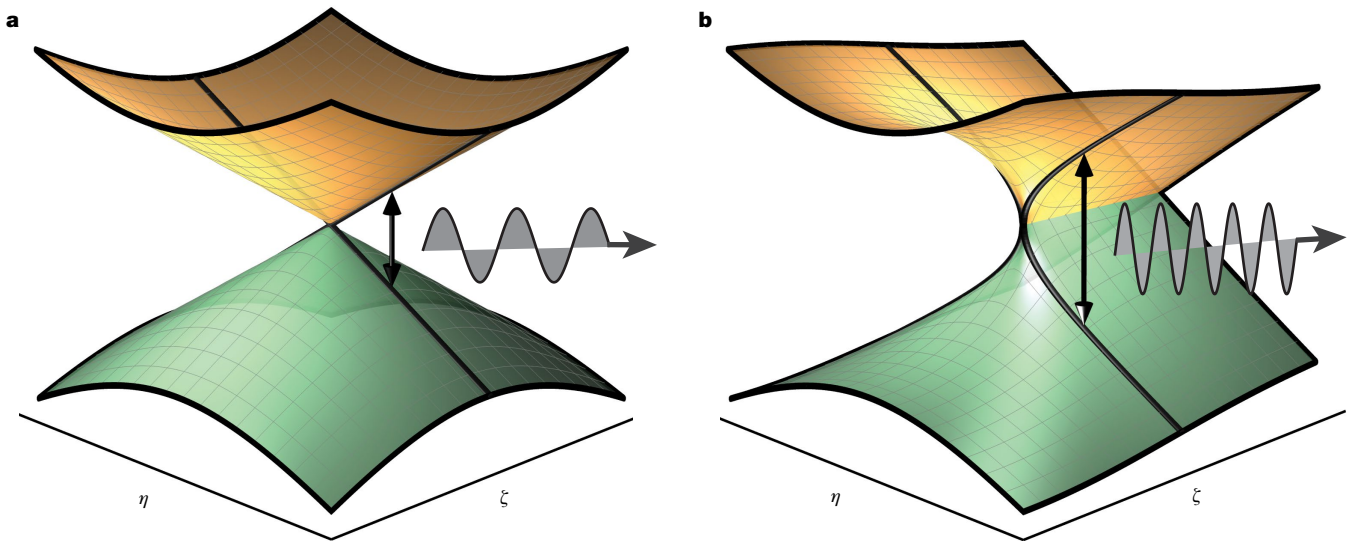
Sensing involves the detection of the signature that a perturbing agent leaves on a system. In optics and many other fields, resonant sensors are made to be as lossless as possible so as to exhibit high quality factors<sup>14–17</sup>. As a result, their response is governed by standard perturbation theory, suited for loss-free or Hermitian arrangements<sup>15</sup>. Recently, however, there has been a growing realization that non-Hermitian systems biased at exceptional points (EPs)<sup>18,19,20</sup>, can react much more drastically to external perturbations<sup>9,10,21</sup>. This EP-enhanced sensitivity—a direct byproduct of Puiseux generalized expansions—is fundamental by nature. In particular, for a system supporting an  $N$ th-order EP, where  $N$  eigenvalues coalesce and their corresponding eigenvectors collapse on each other, the reaction to a perturbation ( $\epsilon$ ) is expected to follow an  $N$ th-root behaviour<sup>9</sup> ( $\epsilon^{1/N}$ ). This is in stark contrast to Hermitian systems, where the sensing response is at best of order  $\epsilon$ . Given that  $\epsilon^{1/N} \gg \epsilon$  for  $|\epsilon| \ll 1$ , this opens up new possibilities for designing ultrasensitive sensors based on such non-Hermitian spectral singularities<sup>9–11</sup>. For illustration purposes, Fig. 1 provides a comparison between the eigenvalue surfaces associated with a Hermitian (Fig. 1a) two-level system ( $N=2$ ) and its corresponding non-Hermitian counterpart (Fig. 1b) when plotted in a two-parameter space around their corresponding spectral degeneracies. As shown in Fig. 1b, the presence of an EP forces the two Riemann manifolds to become strongly intertwined with each other—an attribute that could in turn be used to enhance the performance of a sensor<sup>22</sup>.

Given that sensing is important in many fields, the emerging idea of boosting the sensitivity of a particular system via non-Hermitian degeneracies could have substantial ramifications across several technical areas. Here, we show that the sensitivity of a standard helium–neon (He–Ne) RLG can be drastically enhanced provided that its resonator is judiciously modified so as to support an EP. Figure 2 depicts a

schematic of the non-Hermitian RLG used in this study. As opposed to a standard RLG, the retrofitted cavity involves a Faraday rotator (FR) and a half-wave plate (HWP). These two elements, acting in conjunction with the Brewster windows (BW) incorporated on both ends of the He–Ne gain tube, can be used to introduce a differential loss contrast (or gain contrast),  $\Delta\gamma$ , between the clockwise (CW) and the counter-clockwise (CCW) lasing modes. The method used to achieve this is depicted in Fig. 2a, where the evolution of the state of the polarization associated with the two counter-rotating modes is provided at three consecutive points (A, B, C) in the cavity. In this arrangement, the BWs allow only  $x$ -polarized light to circulate in the cavity while rejecting the  $y$  component. As a result, the CW mode enters the FR as  $x$ -polarized at point A. Because of the magneto-optic effect, the polarization subsequently rotates by a small angle  $\alpha$  (point B). Under the action of the HWP, the angle between the linear electric-field component and the preferred  $x$  axis is  $2\beta - \alpha$  (point C), where the small angle  $\beta$  denotes the orientation of the fast axis of the HWP with respect to the  $x$ – $y$  coordinate system. On the other hand, because of non-reciprocity, although the CCW mode also starts as  $x$ -polarized at point C, it exits at an angle of  $2\beta + \alpha$  with respect to the  $x$  axis (point A) after traversing the same two optical components. Therefore, as clearly indicated in Fig. 2a, the CW mode is expected to experience lower losses than its CCW counterpart does, after passing through the BWs of the He–Ne tube. Hence, a differential loss ( $\Delta\gamma$ ) can be introduced between these two counter-rotating modes. Finally, to establish an EP in this cavity, it is necessary to counteract this differential loss with a mode-coupling process<sup>22</sup>. In our system, the coupling between the CW and CCW modes is readily induced using a weakly scattering object (for example, an etalon), as shown in Fig. 2a. The aforementioned processes can be formally described by employing a Jones calculus approach for the elements

<sup>1</sup>CREOL, The College of Optics and Photonics, University of Central Florida, Orlando, FL, USA. <sup>2</sup>Institute for Theoretical Physics, Vienna University of Technology (TU Wien), Vienna, Austria.

<sup>3</sup>Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA, USA. \*e-mail: khajavikhan@usc.edu



**Fig. 1 | Conceptual illustrations comparing the eigenvalue surfaces associated with Hermitian and non-Hermitian two-level systems. a**, The real part of the eigenvalues plotted in parameter space ( $\eta$ – $\zeta$ ; normalized detuning,  $\eta$ , versus normalized coupling/gain–loss contrast,  $\zeta$ ) when the arrangement is Hermitian. Because of the Hermiticity, this system responds linearly to

perturbations. **b**, The real part of the eigenfrequency surface corresponding to a non-Hermitian configuration in the same parameter space. In the presence of an EP, the two Riemann manifolds are strongly intertwined, leading to a square-root response to perturbations, as indicated by the frequency of the emitted signal. Using this system, an enhanced sensitivity to small changes is expected.

involved (HWP, FR, BW, scattering object), where the polarization state of the CW and CCW waves can be monitored after each pass through the following transfer matrix  $T = S_{SC} \times P \times J_{HWP} \times J_{FR} \times J_{BW}$  (see Supplementary Information). In this expression,  $S_{SC}$  represents a conservative scattering matrix (producing coupling) and  $P$  denotes a phase accumulation matrix that can in principle account for the Sagnac shift<sup>3</sup>. The matrices  $J_{HWP}$ ,  $J_{FR}$  and  $J_{BW}$  are the respective Jones matrices describing the change of polarization after each element<sup>23,24</sup>.

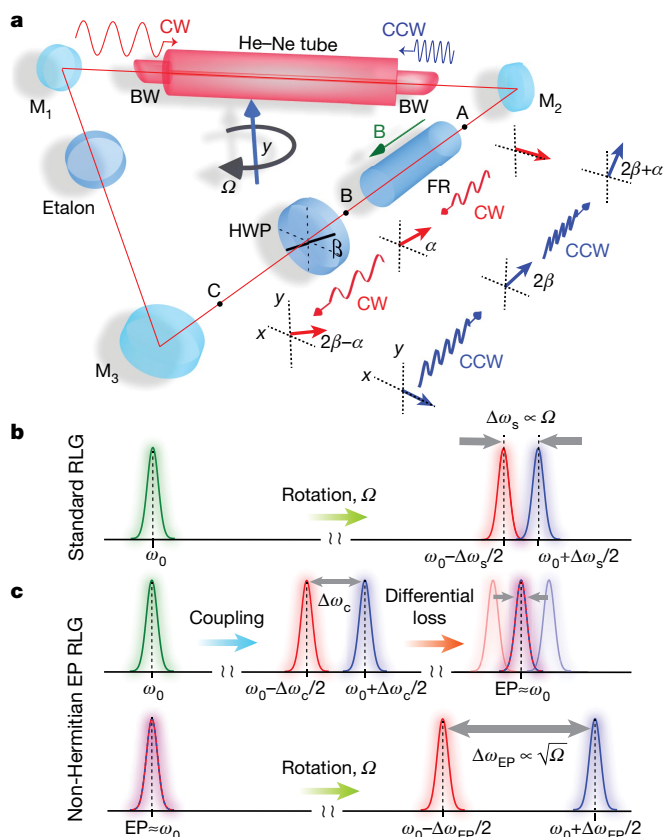
To experimentally demonstrate this enhanced Sagnac sensitivity, we use a custom-made, educational-grade He–Ne RLG (purchased from Luhs; <https://luhs.de/lm-0600-hene-laser-gyroscope.html>). The triangular cavity has a length of 138 cm and supports a free spectral range of about 216 MHz at 632.8 nm. The maximum loss that can be afforded in this system is approximately 3.6%. This resonator is then retrofitted with a terbium gallium garnet (TGG) Faraday element that can provide up to about 4° rotation at a magnetic induction of about 80 mT. This is used in conjunction with a HWP with a rotation angle that can vary in a controlled manner with a resolution of 0.005°. An etalon in the cavity promotes lasing in a specific longitudinal mode while providing some level of coupling between the CW and CCW modes. Other elements, such as the TGG, also contribute to this coupling. Overall, the system is designed to allow maximum tunability in establishing an EP.

Figure 2b, c provides a comparison between the principles of operation of a standard RLG and the EP-based RLG arrangement used in this study. In the former configuration, the Sagnac effect produces a shift ( $\pm\Delta\omega_s/2$ ) in the lasing CW and CCW angular frequencies (which at rest coincide at  $\omega_0$ ), where the beating frequency  $\Delta\omega_s/(2\pi) = 4A\Omega/(\lambda_0 L)$  depends on the angular velocity  $\Omega$  of the rotating frame, the area  $A$  enclosed by the light path (of perimeter  $L$ ) and on the emission wavelength,  $\lambda_0 = 2\pi c/\omega_0$  ( $c$ , speed of light in vacuum). Evidently, the beating frequency  $\Delta\omega_s/(2\pi)$  in this Hermitian setup (which is electronically detected) is always proportional to  $\Omega$  and is dictated by geometrical constraints (Fig. 2b). The situation is entirely different for the non-Hermitian configuration, where the carrier angular frequency  $\omega_0$  can split by  $\pm\Delta\omega_e/2$  even in the absence of rotation because of coupling effects arising from the scatterer (Fig. 2c). In this same static frame, by adjusting the differential loss  $\Delta\gamma$ , these two resonances can fuse

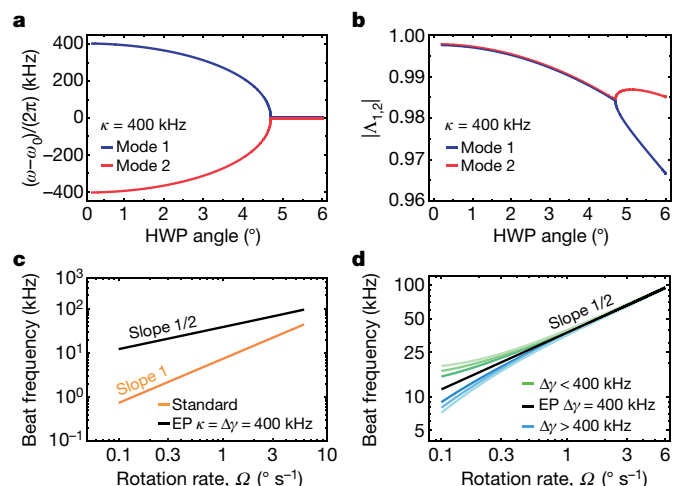
with each other once again at about  $\omega_0$ , thus marking the presence of an EP. After the system is set at an EP, upon rotation  $\Omega$ , the Sagnac shifts  $\pm\Delta\omega_s/2$  induce two new angular frequency lines at  $\omega_0 \pm \Delta\omega_{EP}/2$  (Fig. 2c). In this case, the beating frequency  $\Delta\omega_{EP}/(2\pi)$  is no longer proportional to  $\Omega$ , but instead varies in an enhanced fashion because in this regime  $\Delta\omega_{EP} \propto \sqrt{\Omega}$ , as expected when operating in the vicinity of an EP (Fig. 2c).

The frequency eigenvalues of the non-Hermitian RLG can be directly obtained from the transfer matrix  $T$ , after imposing periodic boundary conditions. From this, the induced non-Hermitian splitting  $\Delta\omega_{EP}$  can be obtained, which interestingly enough remains unaffected even in the presence of gain saturation (see Supplementary Information). On the basis of these results, under rest conditions, one can compute the frequency split associated with the CW and CCW counter-propagating modes in our system as a function of the HWP angle when, for example, the coupling strength is set to  $\kappa = 400$  kHz (Fig. 3a). In this case, an EP appears at  $\beta \approx 4.7^\circ$ . The corresponding magnitude of the complex eigenvalues  $|A_{1,2}|$  of the system is plotted in Fig. 3b. The frequency beating signals expected from the Hermitian (orange line) and the non-Hermitian (black line) configurations of the RLG are plotted in Fig. 3c as a function of  $\Omega$ . In the non-Hermitian case, we assume that the system is positioned at an EP ( $\kappa = \Delta\gamma$ ) when  $\kappa = 400$  kHz. The EP enhancement of the Sagnac shift is evident in this figure. For these parameters, if, for example, the system rotates at  $\Omega = 1^\circ \text{ s}^{-1}$ , the Sagnac signal from the unmodified version of this RLG (Hermitian) is approximately 7.325 kHz, whereas the signal from the retrofitted (EP-based) system is expected to be about 5.2 times larger. Finally, Fig. 3d shows the change of beat note as a function of gyration speed when the system deviates from the EP (by 0.05% to 0.1% of the coupling strength). Although ideally one must keep the system at the EP, for small deviations the resulting error appears to be negligible.

Figure 4a depicts experimental results obtained from our non-Hermitian RLG system when it was biased at an EP. In our experiments, before each set of measurements performed, the system was positioned at an EP by monitoring the beat note as a function of the HWP rotation angle (gain–loss contrast), that is, setting the beat frequency as close as possible to zero. To do so, the HWP rotation angle was adjusted using a motorized rotation stage while the other components in the system



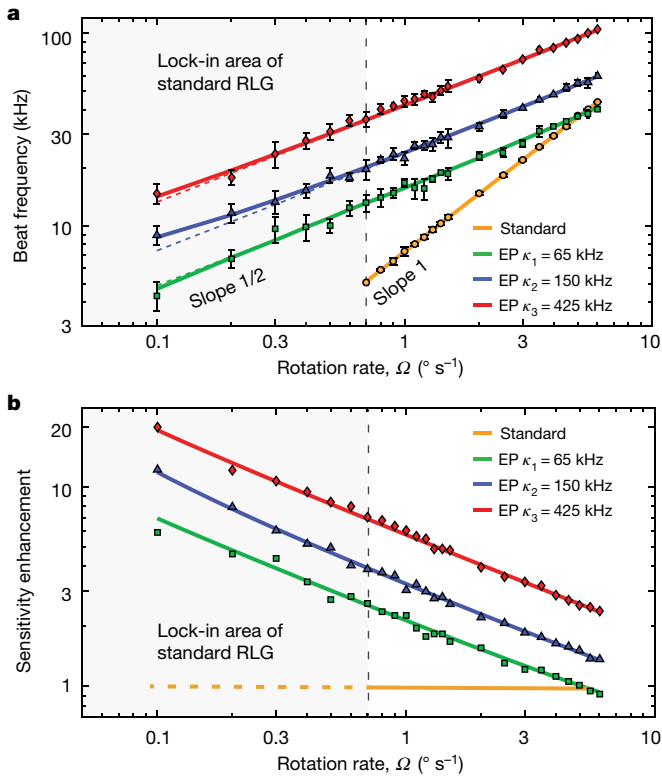
were left intact. The figure provides data corresponding to three different coupling strengths, along with data from the standard unmodified RLG arrangement. These results are plotted in a log-log scale as a function of the rotation rate  $\Omega$  for  $\kappa = 65$  kHz, 150 kHz, 425 kHz. Whereas the response of the standard configuration is linear with respect to  $\Omega$  (slope of 1), the same is not true for its non-Hermitian embodiment. In the latter case, the response is found to vary as the square root of the rotation rate  $\Omega$ , as is evident from the slope of the accompanying three curves, which is very close to  $1/2$ —a clear indication that an EP is at play. Our experimental observations clearly show that the scale factor of the Sagnac effect is substantially boosted by exploiting the very properties of EPs. The resulting Sagnac enhancement factors (with respect to the standard arrangement) are plotted in Fig. 4b for the same three cases. For  $\kappa = 425$  kHz, a sensitivity boost of more than an order



of magnitude is observed when  $\Omega = 0.4^\circ \text{ s}^{-1}$ . The reported minimum gyration speed,  $\Omega = 0.1^\circ \text{ s}^{-1}$ , is imposed by the limited rotation capability of the apparatus. The estimated rotation rate is obtained from the beat frequency by applying the transfer functions associated with the Hermitian and non-Hermitian arrangements. These transfer functions are illustrated in Fig. 5a, where it can be observed that for small angular velocities, not only the absolute value of the beat frequency ( $\Delta v_{EP} > \Delta v_S$ ; where  $\Delta v_{EP} = \Delta\omega_{EP}/(2\pi)$  and  $\Delta v_S = \Delta\omega_S/(2\pi)$  is the rotation-induced beat frequency in the EP-based RLG and the standard RLG, respectively) increases dramatically for the EP-based system, but also an incremental step in the rotation rate is transferred to a much larger difference in the beat frequency ( $|\Delta v_{EP,2} - \Delta v_{EP,1}| > |\Delta v_{S,2} - \Delta v_{S,1}|$ ). As a result, the resolution of the estimated rotation speed is potentially improved. Figure 5b, c displays the error bars on the estimated rotation rates, as obtained experimentally for the Hermitian and non-Hermitian arrangements, respectively. In a standard gyroscope, the relationship between the applied and estimated angular velocities is linear. On the other hand, owing to the nonlinear transfer function associated with the non-Hermitian system, these two quantities are not on an equal footing anymore. In this respect, only when considering the nonlinearity of the transfer function, the errors on the estimated rotation rate can be interpreted correctly. Consequently, at higher rotation speeds, the modified gyroscope displays larger error bars, as shown in Fig. 5c.

72 | Nature | Vol 576 | 5 December 2019

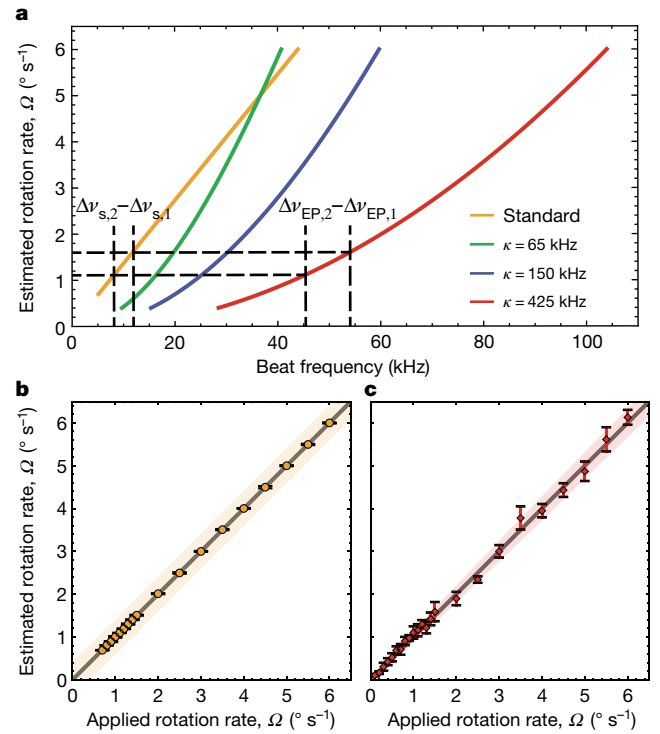




**Fig. 4 | Measured beat frequency and sensitivity enhancement factor versus rotation rate.** **a**, Beat frequency versus rotation rate  $\Omega$  (in log–log scale) for a standard RLG (orange data marks) and a non-Hermitian RLG at three different coupling strengths,  $\kappa_1 = 65$  kHz (green),  $\kappa_2 = 150$  kHz (blue) and  $\kappa_3 = 425$  kHz (red). The dashed lines correspond to theoretical calculations in which the non-Hermitian system is biased exactly at the EP ( $\Delta\gamma_i = \kappa_i$ ). The solid lines represent fitted data obtained when the system is slightly detuned from the EP (see Supplementary Information, equation (11)) (here  $\Delta\gamma_1 = 1.0003\kappa_1$ ,  $\Delta\gamma_2 = 0.9992\kappa_2$ ,  $\Delta\gamma_3 = 0.9999\kappa_3$ ). The orange line has a slope of unity, indicating that the Sagnac shift in the standard cavity varies linearly with  $\Omega$ . By contrast, the slope associated with the non-Hermitian curves is approximately  $1/2$ , indicating the presence of an EP. Moreover, whereas in the standard RLG the lock-in effect limits gyration measurements below  $\Omega = 0.7^\circ \text{ s}^{-1}$  (shaded region), the EP-based configuration is capable of detecting smaller rotation rates (only limited by the resolution of the step motor,  $0.1^\circ \text{ s}^{-1}$ ). The error bars show one standard deviation from the mean of the collected data. **b**, The sensitivity enhancement, defined as the ratio of the non-Hermitian beat frequency to that of a standard RLG, is obtained from the measured data for the aforementioned three coupling strengths. For  $\Omega < 0.7^\circ \text{ s}^{-1}$ , the sensitivity enhancement is calculated using the anticipated value of the beat frequency from the standard RLG, provided that lock-in does not occur. The solid lines (red, blue and green) represent theoretical curves corresponding to the parameters used in our experiments.

The advantage of the EP-based gyroscope becomes apparent at smaller velocities, where the error in the estimated rotation rates decreases rapidly. This is depicted in Fig. 5b, c, where a noise component of 3 kHz has been added to the ideal system (shown as orange and red shaded regions).

Several factors must be considered when using non-Hermitian arrangements for sensing purposes. First and foremost is appreciating the difference between sensitivity and detection limit<sup>25</sup>. In non-Hermitian settings, the sensitivity enhancement is a fundamental feature that is dictated by mathematical properties, governed by the perturbation expansion around an EP. The detection limit, on the other hand, depends on the physical system and is primarily determined by the net gain (or loss), as well as the correlation between the laser noise associated with the two resonances<sup>26,27</sup>. In this regard, one in principle



**Fig. 5 | Transfer functions and estimated rotation rates.** **a**, The transfer function of the standard system (orange line)—that is, its response to a beat frequency—is compared to that of the EP-based RLG for  $\kappa = 65, 150, 425$  kHz (green, blue and red lines, respectively). For the non-Hermitian gyroscope, at small rotation rates both the absolute values of the beat frequency ( $\Delta\nu_{\text{EP}} > \Delta\nu_s$ ) and of the beat frequency differences ( $|\Delta\nu_{\text{EP},2} - \Delta\nu_{\text{EP},1}| > |\Delta\nu_{s,2} - \Delta\nu_{s,1}|$ ) for an incremental step in the rotation rate increase dramatically. **b**, **c**, Predicted rotation rate for the standard RLG (**b**) and for the modified non-Hermitian RLG at  $\kappa = 425$  kHz (**c**), obtained by applying the associated transfer functions to the measured beat frequencies. The shaded areas demonstrate the effect of noise (3 kHz) on the estimated rotation rates. The error bars show one standard deviation from the mean of the collected data.

can increase the net gain while keeping the RLG at the EP by managing the gain contrast to boost both the sensitivity and the detection limit—as we did in our design. As expected from the Schawlow–Townes formula, an increase in the net (average) gain of the system (or the output power) will reduce the linewidth of the laser. This in turn tends to compensate for the linewidth broadening near the EP while allowing one to exploit the larger sensitivity afforded by such singularities. Another technical issue is how closely one can reach and stabilize the system at the EP<sup>28,29</sup>. In our experiment, we fully rely on positioning the RLG at the EP before each set of measurements, by visually monitoring the beat note as a function of the HWP rotation angle (gain–loss contrast). In future devices to be used in the field, one may need to actively control the system to remain biased at the EP. Such approaches have been suggested elsewhere<sup>11,30</sup>.

In conclusion, we have demonstrated for the first time, to our knowledge, a new class of non-Hermitian RLGs that can display an enhanced Sagnac sensitivity. This is accomplished by exploiting the intriguing properties of a special family of non-Hermitian spectral singularities, the EPs. At these points, the RLG response has a square-root dependence on the gyration speed, in contrast to the linear response observed in standard arrangements. The proposed configuration may inspire new technological developments in various settings in which measuring low rotation rates via ultracompact systems is highly attractive. Finally, the idea of transforming a standard measuring apparatus into an EP-based device with superior sensitivity may have important ramifications in other areas of science and technology.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1780-4>.

- Armenise, M. N., Ciminelli, C., Dell'Olio, F. & Passaro, V. M. *Advances in Gyroscope Technologies* (Springer, 2010).
- Post, E. J. Sagnac effect. *Rev. Mod. Phys.* **39**, 475–493 (1967).
- Chow, W. W. et al. The ring laser gyro. *Rev. Mod. Phys.* **57**, 61–104 (1985).
- Macek, W. M. & Davis, D. T. M. Jr Rotation rate sensing with traveling-wave ring lasers. *Appl. Phys. Lett.* **2**, 67–68 (1963).
- Boyd, R. W. Slow and fast light: fundamentals and applications. *J. Mod. Opt.* **56**, 1908–1915 (2009).
- Shahriar, M. S. et al. Ultrahigh enhancement in absolute and relative rotation sensing using fast and slow light. *Phys. Rev. A* **75**, 053807 (2007).
- Smith, D. D., Chang, H., Arissian, L. & Diels, J. C. Dispersion-enhanced laser gyroscope. *Phys. Rev. A* **78**, 053824 (2008).
- Kaplan, A. E. & Meystre, P. Enhancement of the Sagnac effect due to nonlinearly induced nonreciprocity. *Opt. Lett.* **6**, 590–592 (1981).
- Hodaiei, H. et al. Enhanced sensitivity at higher-order exceptional points. *Nature* **548**, 187–191 (2017); erratum 551, 658–191 (2017).
- Chen, W., Özdemir, Ş. K., Zhao, G., Wiersig, J. & Yang, L. Exceptional points enhance sensing in an optical microcavity. *Nature* **548**, 192–196 (2017).
- Ren, J. et al. Ultrasensitive micro-scale parity-time-symmetric ring laser gyroscope. *Opt. Lett.* **42**, 1556–1559 (2017).
- Sunada, S. Large Sagnac frequency splitting in a ring resonator operating at an exceptional point. *Phys. Rev. A* **96**, 033842 (2017).
- Grant, M. J. & Diggonnet, M. J. F. Loss-gain coupled ring resonator gyroscope. In *Proc. SPIE Optical, Opto-Atomic, and Entanglement-Enhanced Precision Metrology* Vol. 10934 (SPIE, 2019).
- Vahala, K. J. Optical microcavities. *Nature* **424**, 839–846 (2003).
- Vollmer, F. & Arnold, S. Whispering-gallery-mode biosensing: label-free detection down to single molecules. *Nat. Methods* **5**, 591–596 (2008).
- Lu, T. et al. High sensitivity nanoparticle detection using optical microcavities. *Proc. Natl Acad. Sci. USA* **108**, 5976–5979 (2011).
- Liang, W. et al. Resonant microphotonic gyroscope. *Optica* **4**, 114–117 (2017).
- Makris, K. G., El-Ganainy, R., Christodoulides, D. N. & Musslimani, Z. H. Beam dynamics in PT symmetric optical lattices. *Phys. Rev. Lett.* **100**, 103904 (2008).
- Klaiman, S., Gunther, U. & Moiseyev, N. Visualization of branch points in PT-symmetric waveguides. *Phys. Rev. Lett.* **101**, 080402 (2008).
- Moiseyev, N. *Non-Hermitian Quantum Mechanics* (Cambridge Univ. Press, 2011).
- Wiersig, J. Enhancing the sensitivity of frequency and energy splitting detection by using exceptional points: application to microcavity sensors for single-particle detection. *Phys. Rev. Lett.* **112**, 203901 (2014).
- El-Ganainy, R. et al. Non-Hermitian physics and PT symmetry. *Nat. Phys.* **14**, 11–19 (2018).
- Yariv, A. & Yeh, P. *Photonics: Optical Electronics in Modern Communications* (Oxford Univ. Press, 2006).
- Khajavikhan, M., John, K. & Leger, J. R. Experimental measurements of supermodes in superposition architectures for coherent laser beam combining. *IEEE J. Quantum Electron.* **46**, 1221–1231 (2010).
- Hu, J., Sun, X., Agarwal, A. & Kimerling, L. C. Design guidelines for optical resonator biochemical sensors. *J. Opt. Soc. Am. B* **26**, 1032–1041 (2009).
- LIGO Scientific and Virgo Collaboration. GW170104: observation of a 50-solar-mass binary black hole coalescence at redshift 0.2. *Phys. Rev. Lett.* **118**, 221101 (2017).
- Collett, M. J., Loudon, R. & Gardiner, C. W. Quantum theory of optical homodyne and heterodyne detection. *J. Mod. Opt.* **34**, 881–902 (1987).
- Mortensen, N. A. et al. Fluctuations and noise-limited sensing near the exceptional point of parity-time-symmetric resonator systems. *Optica* **5**, 1342–1346 (2018).
- Zhang, M. et al. Quantum noise theory of exceptional point amplifying sensors. *Phys. Rev. Lett.* **123**, 180501 (2019).
- De Carlo, M., De Leonardis, F. & Passaro, V. M. Design rules of a microscale PT-symmetric optical gyroscope using group IV platform. *J. Light. Technol.* **36**, 3261–3268 (2018).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

## Data availability

All data that support the findings of this study are available within the paper and the Supplementary Information and are available from the corresponding author upon reasonable request.

**Acknowledgements** We acknowledge support from the US Air Force Office of Scientific Research (FA9550-14-1-0037), Office of Naval Research (N00014-16-1-2640, N00014-18-1-2347, N00014-19-1-2052), National Science Foundation (ECCS1454531, DMR-1420620, ECCS1757025), Army Research Office (W911NF-16-1-0013, W911NF-17-1-0481), US-Israel Binational Science Foundation (BSF) (2016381), DARPA (D18AP00058, HR00111820042, HR00111820038) and the European Commission Project 'Non-Hermitian Quantum Wave Engineering' (NHQWAVE, MSCA-RISE 691209). We thank W. Luhs for help in setting up the

gyroscope and for performing some of the initial measurements, S. Milady, S. Rotter and K. Vahala for technical discussions, and S. Rotter for supporting this project through funding for A.S.

**Author contributions** All authors contributed equally to this work.

**Competing interests** The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1780-4>.

**Correspondence and requests for materials** should be addressed to M.K.

**Peer review information** *Nature* thanks Chia Wei Hsu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



# Visualizing Poiseuille flow of hydrodynamic electrons

<https://doi.org/10.1038/s41586-019-1788-9>

Received: 26 May 2019

Accepted: 11 October 2019

Published online: 4 December 2019

Joseph A. Sulpizio<sup>1,8</sup>, Lior Ella<sup>1,8</sup>, Asaf Rozen<sup>1,8</sup>, John Birkbeck<sup>2,3</sup>, David J. Perello<sup>2,3</sup>, Debarghya Dutta<sup>1</sup>, Moshe Ben-Shalom<sup>2,3,4</sup>, Takashi Taniguchi<sup>5</sup>, Kenji Watanabe<sup>5</sup>, Tobias Holder<sup>1</sup>, Raquel Queiroz<sup>1</sup>, Alessandro Principi<sup>2</sup>, Ady Stern<sup>1</sup>, Thomas Scaffidi<sup>6,7</sup>, Andre K. Geim<sup>2,3</sup> & Shahal Ilani<sup>1\*</sup>

Hydrodynamics, which generally describes the flow of a fluid, is expected to hold even for fundamental particles such as electrons when inter-particle interactions dominate<sup>1</sup>. Although various aspects of electron hydrodynamics have been revealed in recent experiments<sup>2–11</sup>, the fundamental spatial structure of hydrodynamic electrons—the Poiseuille flow profile—has remained elusive. Here we provide direct imaging of the Poiseuille flow of an electronic fluid, as well as a visualization of its evolution from ballistic flow. Using a scanning carbon nanotube single-electron transistor<sup>12</sup>, we image the Hall voltage of electronic flow through channels of high-mobility graphene. We find that the profile of the Hall field across the channel is a key physical quantity for distinguishing ballistic from hydrodynamic flow. We image the transition from flat, ballistic field profiles at low temperatures into parabolic field profiles at elevated temperatures, which is the hallmark of Poiseuille flow. The curvature of the imaged profiles is qualitatively reproduced by Boltzmann calculations, which allow us to create a ‘phase diagram’ that characterizes the electron flow regimes. Our results provide direct confirmation of Poiseuille flow in the solid state, and enable exploration of the rich physics of interacting electrons in real space.

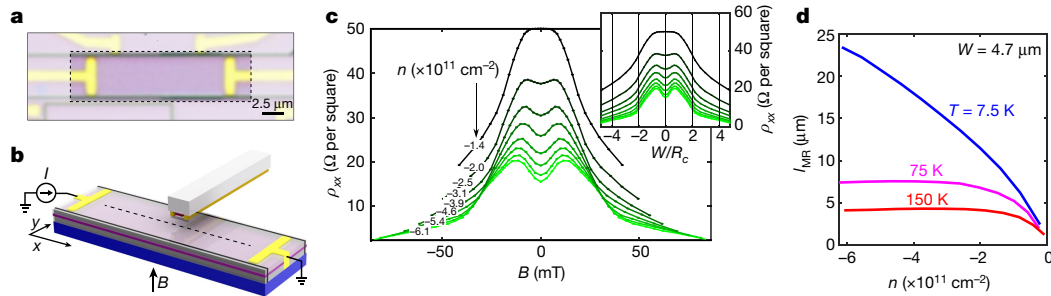
The notion of viscosity arises in hydrodynamics to describe the diffusion of momentum in a fluid under the application of shear stress. When scattering between constituent fluid particles becomes dominant, viscosity manifests as an effective frictional force between fluid layers. The hallmark of such hydrodynamic transport in a channel is a parabolic, or Poiseuille, velocity flow profile, which typifies familiar phenomena like water flowing through a pipe. Electron flow has long been predicted<sup>1</sup> to undergo hydrodynamic transport when the rate of momentum-conserving Coulomb scattering between electrons exceeds that of momentum-relaxing scattering from impurities, boundaries and phonons<sup>2,13,14</sup>. The implications of a dominant viscous force on electronic flow have been studied in a range of theoretical work<sup>15–21</sup>. While initial efforts were based on linearized Navier–Stokes equations, describing electron hydrodynamics in the context of diffusive transport<sup>14,18–20,22</sup>, there is an evolving understanding that a central part of the physics is the emergence of hydrodynamics from ballistic flow<sup>3,6,7,10,23–28</sup>. Reaching the hydrodynamic regime experimentally requires materials of high purity so that Ohmic transport can be minimized, which is now possible in a growing number of high-mobility systems. Indeed, recent experiments have demonstrated the existence of negative non-local resistance<sup>3,7</sup>, superballistic flow<sup>6</sup>, signatures of Hall viscosity<sup>9,11</sup>, breakdown of the Wiedemann–Franz law<sup>4,8</sup>, and anomalous scaling of resistance

with channel width<sup>5</sup>, all phenomena associated with hydrodynamic electron flow. Yet the direct observation of the fundamental Poiseuille flow profile has remained elusive.

In this work, we provide the first, to our knowledge, spatial imaging of Poiseuille flow of hydrodynamic electrons, as well as the evolution from ballistic to hydrodynamic flow. We use a scanning carbon nanotube single-electron transistor (SET) to non-invasively image maps of the longitudinal and Hall voltage of electrons flowing through high-mobility graphene/hexagonal boron nitride (hBN) channels<sup>12</sup>. By varying the carrier density (degenerate regime away from charge neutrality) and temperature, we tune the two relevant length scales controlling electron flow: the momentum-relaxing mean free path, set by electron-impurity and electron-phonon scattering, and the momentum conserving mean free path, set by electron–electron interactions. We find that the spatial profile of the Hall field across the channel is key to distinguishing the evolution from ballistic into hydrodynamic flow. At low temperatures, we observe flat profiles associated with ballistic flow. At higher temperatures the profiles become parabolic, with curvature approaching that of ideal Poiseuille flow. Overall, we find that Boltzmann kinetic equations qualitatively reproduce our observations, although at the highest temperatures they underestimate the curvature of the Hall field profiles. Finally, we show that this

<sup>1</sup>Department of Condensed Matter Physics, Weizmann Institute of Science, Rehovot, Israel. <sup>2</sup>School of Physics and Astronomy, University of Manchester, Manchester, UK. <sup>3</sup>National Graphene Institute, University of Manchester, Manchester, UK. <sup>4</sup>Department of Physics and Astronomy, Tel-Aviv University, Tel Aviv, Israel. <sup>5</sup>National Institute for Materials Science, Tsukuba, Japan.

<sup>6</sup>Department of Physics, University of California, Berkeley, CA, USA. <sup>7</sup>Department of Physics, University of Toronto, Toronto, Ontario, Canada. <sup>8</sup>These authors contributed equally: J. A. Sulpizio, L. Ella, A. Rozen. \*e-mail: shahal.ilani@weizmann.ac.il



**Fig. 1 | Overview of graphene channel device and imaging of magnetoresistance.** **a**, Optical image of graphene channel device used for imaging electron flow, consisting of a high-mobility monolayer of graphene sandwiched between hBN layers (purple) and electrical contact electrodes (yellow) on top of the conducting Si/SiO<sub>2</sub> back gate. The dark lines are etched walls that define a channel of width  $W = 4.7 \mu\text{m}$  and length  $L = 15 \mu\text{m}$  (outlined with dashed box; scale bar,  $2.5 \mu\text{m}$ ). **b**, Rendering of scanning SET imaging performed in experiments. The nanotube-based SET is positioned at the end of a scanning probe cantilever, and is rastered across the channel (graphene in purple, sandwiched between hBN layers atop a Si/SiO<sub>2</sub> substrate in blue) to locally image the potential generated by the electrical current  $I$  in a perpendicular magnetic field  $B$ . **c**, Magnetoresistance of graphene channel at a temperature of  $T = 7.5 \text{ K}$ , antisymmetrized in  $B$ , imaged non-invasively with

scanning SET. The SET is scanned along the centreline of the channel (black dashed line in **b**) to image the potential drop  $\Delta\phi$  in order to extract the longitudinal resistance  $\rho_{xx} = W \frac{\Delta\phi}{\Delta x} / I$  as a function of magnetic field  $B$  for different charge carrier densities  $n$  (black curve is low density, high density in green; numbers label the density of each curve). Inset, the same  $\rho_{xx}$  data plotted as a function of  $W/R_c$ , which is proportional to  $B$  (see text). At high density, the magnetoresistance curves show a double-peaked structure, indicating ballistic transport with diffusive walls (see Methods). **d**, Momentum-relaxing mean free path  $l_{MR}$  in the bulk of the graphene channel as a function of carrier density for several temperatures. The SET is maintained at liquid helium temperature throughout all measurements<sup>11</sup>. The value of  $l_{MR}$  is deduced from  $\rho_{xx}(B)$  and is described in Methods, which also presents the associated mobility (Extended Data Fig. 1).

curvature is the distinctive metric for characterizing the different flow regimes, allowing us to construct a phase diagram and map the regions explored by the experiment.

The devices we studied are high-mobility monolayer graphene/hBN heterostructures patterned into channels of various lengths,  $L$ , and widths,  $W$ . Below we present data from a device with  $W = 4.7 \mu\text{m}$  and  $L = 15 \mu\text{m}$  (Fig. 1a), but similar results have been obtained for a device with a different width, aspect ratio, and etched boundaries (see Methods and Extended Data Fig. 5).

We first perform the scanning analogue<sup>12</sup> of transport measurements of longitudinal resistivity,  $\rho_{xx}$ . Flowing current  $I$  through the channel and imaging the potential produced by the flowing electrons,  $\phi(x)$ , along the centreline (dashed line in Fig. 1b) yields  $\rho_{xx} = W \frac{d\phi}{dx} / I$ . Figure 1c shows  $\rho_{xx}$  as a function of perpendicular magnetic field,  $B$ , for various carrier densities,  $n$ , at a temperature of  $T = 7.5 \text{ K}$ . Notably, with increasing  $|n|$ ,  $\rho_{xx}$  evolves from a single- to double-peaked structure. This is a well known signature of ballistic electron transport ( $l_{MR} > W$ , where  $l_{MR}$  is the momentum-relaxing mean free path), when scattering at the walls is diffusive<sup>5,29</sup> (see Methods and Extended Data Fig. 2). The  $B$  dependence of  $\rho_{xx}$  is set by the ratio of  $W$  and the cyclotron radius,  $R_c = \frac{\hbar \sqrt{\pi |n|}}{eB}$  ( $\hbar$  is the reduced Planck constant and  $e$  is the electron charge). For  $|W/R_c| > 2$ , backscattering is strongly suppressed, and Boltzmann theory predicts<sup>24</sup> that  $\rho_{xx}$  is determined primarily by bulk scattering (with correction proportional to  $|W/R_c|^{-1}$ ), allowing us to estimate  $l_{MR}$  (see Methods and Extended Data Fig. 1). Figure 1d plots the extracted  $l_{MR}$  as a function of  $n$  at several different temperatures. For  $T = 7.5 \text{ K}$ ,  $l_{MR}$  exhibits the expected  $|n|$ -dependence, while at  $T = 75 \text{ K}$  and  $150 \text{ K}$ ,  $l_{MR}$  displays a characteristic flat density dependence due to the addition of phonon scattering<sup>30</sup>.

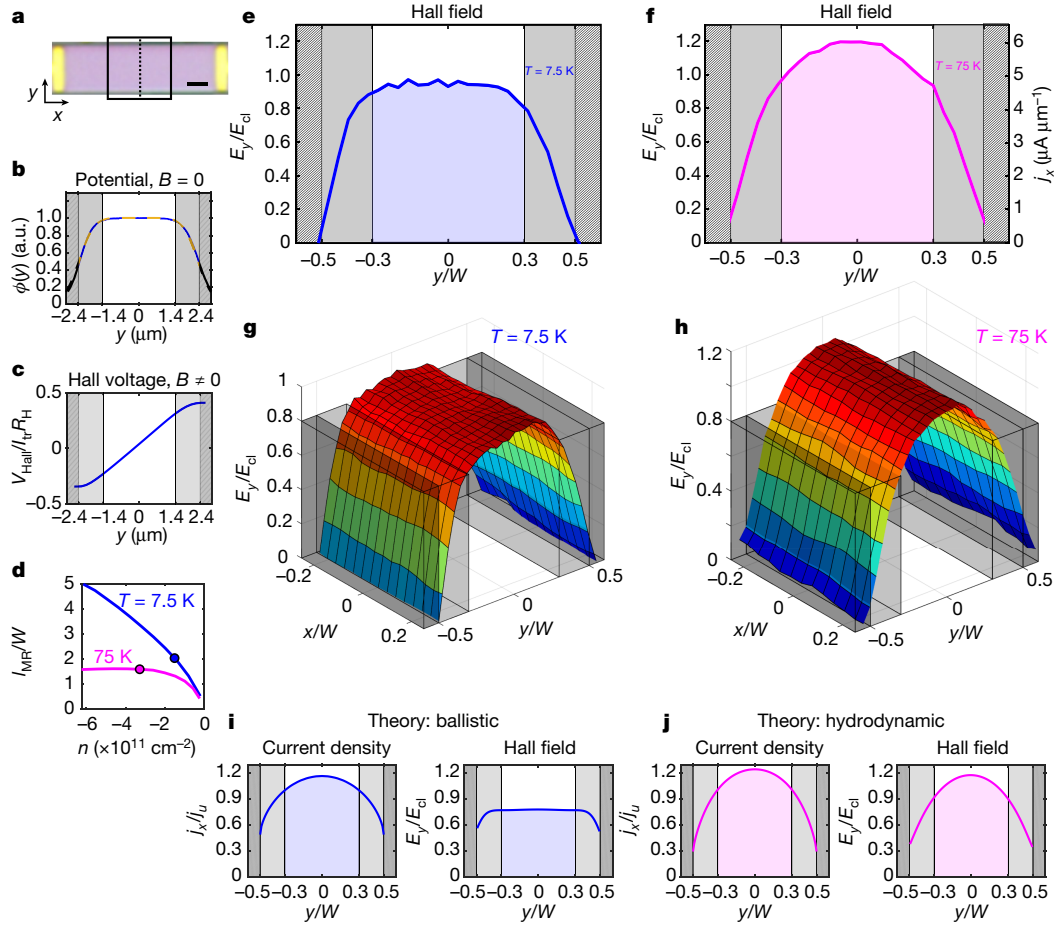
We now turn to the Hall voltage profiles, which are fundamentally related to the current flow profiles of electrons in the channel. We restrict all subsequent analysis to the bulk of the channel ( $|y/W| < 0.3$ ), in which the convolution of the potential jump near the channel edge with the point spread function of our SET due to the imaging height ( $h \approx 880 \text{ nm}$ ) is negligible (Fig. 2a, b). In the Ohmic regime ( $l_{MR} \ll W$ ), there is a local relation between the  $y$ -component of the Hall field,  $E_y = dV_{Hall}/dy$  (where  $V_{Hall}$  is the local Hall voltage), and the current density parallel to the channel axis,  $j_x$ , given by  $E_y = (B/ne)j_x$ . In the hydrodynamic regime, where  $l_{ee} < W$  (where  $l_{ee}$  is the electron–electron scattering

length), the current density is predicted to be parabolic, leading to an analogous relation<sup>31</sup> (see Methods):

$$E_y = \frac{B}{ne} \left( j_x + \frac{1}{2} l_{ee}^2 \partial_y^2 j_x \right) \quad (1)$$

Deep in the hydrodynamic regime, where  $l_{ee}/W \ll 1$ , the local relation between  $E_y$  and  $j_x$  is recovered to a good approximation. Imaging  $E_y(y)$  in these regimes therefore effectively images the current distribution,  $j_x(y)$ . In the ballistic regime, this relation breaks down, leading to a fundamentally different  $E_y$  profile. As we show,  $E_y$  is then a key observable for distinguishing between ballistic and hydrodynamic flows. Figure 2c shows the potential along  $y$  measured at small magnetic fields  $B = \pm 12.5 \text{ mT}$ , antisymmetrized in  $B$ , to yield the Hall voltage profile  $V_{Hall}(y) = \frac{1}{2} [\phi(y, B) - \phi(y, -B)]$ , where  $T = 7.5 \text{ K}$  and  $n = -1.5 \times 10^{11} \text{ cm}^{-2}$ . Note that  $B$  is small enough that the flow remains semiclassical (Landau level filling factor  $\nu \gg 100$  and  $\hbar\omega_c \ll k_B T$ , where  $\omega_c$  is the cyclotron frequency and  $k_B$  is the Boltzmann constant). The  $E_y(y)$  profiles are obtained by numerically differentiating the imaged  $V_{Hall}(y)$  profiles.

We now observe how electron–electron interactions affect the Hall field profiles by comparing imaging at different temperatures:  $T = 7.5 \text{ K}$  in Fig. 2e, and  $T = 75 \text{ K}$  in Fig. 2f. While increased temperature should increase the electron–electron scattering rate (decrease  $l_{ee}$ ) it also increases electron–phonon scattering (decreases the electron–phonon mean free path,  $l_{ph}$ ) and correspondingly reduces  $l_{MR} = (l_{imp}^{-1} + l_{ph}^{-1})^{-1}$ , where  $l_{imp}$  is the impurity scattering mean free path. To best isolate the influence of  $l_{ee}$ , we therefore maintain a nearly constant  $l_{MR}$  across the different temperatures by tuning the carrier density between the measurements (circles in Fig. 2d; see legend for details). Notably, the imaged profile at  $T = 7.5 \text{ K}$  is flat across the bulk of the channel (Fig. 2e). In contrast, the profile at  $T = 75 \text{ K}$  is strongly parabolic (Fig. 2f). The dramatic difference in curvature between these profiles becomes more apparent when we image the full two-dimensional maps of the Hall field (within the black box in Fig. 2a), demonstrating that the profiles are independent of position along the channel (Fig. 2g, h). All measurements are performed at small enough magnetic field ( $W/R_c = 1.3$ ) to minimally influence the profiles, as well as low voltage bias across the channel to avoid electron heating (see Methods and Extended Data Figs. 3, 4).



**Fig. 2 | Imaging ballistic and Poiseuille electron flow profiles.** **a**, Graphene channel with overlay indicating the region over which flow profiles are imaged. One-dimensional profiles are taken along the dashed line and two-dimensional profiles are imaged across the region enclosed by the black box (scale bar, 2.5  $\mu\text{m}$ ). **b**, Potential of flowing electrons,  $\phi$ , as a function of the  $y$  coordinate (dashed line in **a**) imaged at  $B = 0$  (blue curve,  $T = 7.5$  K). The dashed yellow curve is a boxcar function convolved with the point spread function of our SET measurement, determined primarily by the height of our SET detector above the graphene during the scan. Grey-shaded regions ( $0.3 < |y|/W < 0.5$ ) indicate where the smearing of the steps at the edges due to the finite spatial resolution has a non-negligible contribution. **c**, Imaged Hall voltage,  $V_{\text{Hall}}$ , from antisymmetrizing measurements taken at field  $B = \pm 12.5$  mT,  $n = -1.5 \times 10^{11} \text{ cm}^{-2}$  and  $T = 7.5$  K. Normalization  $I_H R_H = 470 \mu\text{V}$ . **d**,  $I_{\text{MR}}$  from Fig. 1d, but now normalized by  $W$ . Dots indicate the carrier densities of the profile imaging in all subsequent panels, where  $n = -1.5 \times 10^{11} \text{ cm}^{-2}$  at 7.5 K and  $n = -3.1 \times 10^{11} \text{ cm}^{-2}$  at

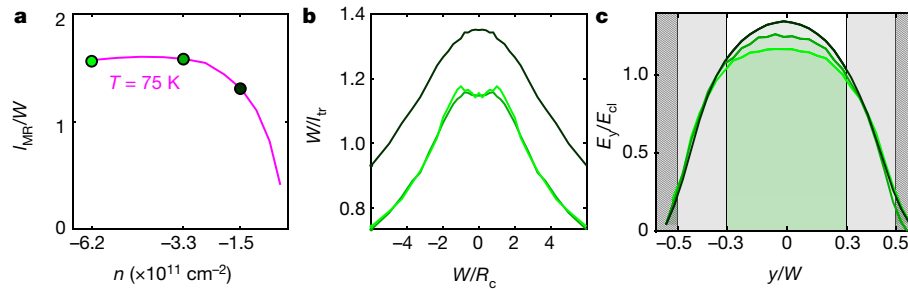
75 K, chosen such that  $I_{\text{MR}}$  is nearly equal for both temperatures. **e**, The Hall field,  $E_y$ , at  $T = 7.5$  K, from measurements at  $B = \pm 12.5$  mT, obtained by numerical differentiation of  $V_{\text{Hall}}$  with respect to  $y$ , normalized by the classical value  $E_{\text{cl}} = (B/ne)l/W = 91 \text{ V m}^{-1}$ . **f**,  $E_y$  at  $T = 75$  K, from measurements at  $B = \pm 18.0$  mT, with  $E_{\text{cl}} = 162 \text{ V m}^{-1}$ . The right  $y$  axis converts the field to units of current density by scaling with  $ne/B$ . **g**, Two-dimensional map of  $E_y$  taken over the boxed region in **a** at  $T = 7.5$  K. **h**, Two-dimensional map of  $E_y$  at  $T = 75$  K. **i, j**, Calculation of the current density  $j_x$  (normalized by  $j_0 = I/W = 2 \text{ A m}^{-1}$  in **i** and  $5.4 \text{ A m}^{-1}$  in **j**), and the Hall field  $E_y/E_{\text{cl}}$  based on the Boltzmann theory with values of  $l_{\text{MR}}$  and  $l_{\text{ee}}$  corresponding to the experimental data in **e** and **f**. In **i**, the values used are  $l_{\text{MR}}/W = 2$  and  $l_{\text{ee}}/W = 4.3$ , whereas for **j**,  $l_{\text{MR}}/W = 1.4$  and  $l_{\text{ee}}/W = 0.16$ . The calculated profiles are convolved with the point spread function of the SET for direct comparison with the experiment. The current density appears parabolic in both the hydrodynamic and ballistic regimes, whereas the  $E_y$  profile is relatively flat in the ballistic regime and parabolic in the hydrodynamic regime.

One naively expects the current density profile,  $j_x(y)$ , to be flat for ballistic flow and parabolic for hydrodynamic Poiseuille flow. However, a full Boltzmann theoretical calculation of the profiles of  $j_x$  and  $E_y$  including  $l_{\text{MR}}$  (Fig. 2i, j and Methods) reveals that this is not the case. The  $j_x$  profile, even deep in the ballistic regime ( $l_{\text{MR}}/W \gg 1$ ), is not flat (see Methods and Extended Data Fig. 8). Figure 2i plots the  $j_x$  profile calculated for  $l_{\text{MR}}/W = 2$  and  $l_{\text{ee}}/W = 4.3$ , consistent with our measurements at  $T = 7.5$  K, showing that  $j_x$  has large curvature. In fact, the Boltzmann theory predicts a strongly curved  $j_x$  profile even for much larger  $l_{\text{MR}}/W$ , showing that the ballistic  $j_x$  profile is not qualitatively different from its hydrodynamic counterpart (an example calculated for  $l_{\text{MR}}/W = 1.4$  and  $l_{\text{ee}}/W = 0.16$  is shown in Fig. 2j), and is therefore a weak marker for the emergence of electron hydrodynamics. In contrast, the Boltzmann theory shows that the  $E_y$  profile differs markedly between ballistic and hydrodynamic flows, making it a way of distinguishing these regimes. In the ballistic regime  $E_y$  is flat (Fig. 2i), and can even

become negatively curved if  $l_{\text{MR}}/W$  is increased further, while in the Poiseuille regime  $E_y$  is positively curved (Fig. 2j).

The  $E_y$  profile in Fig. 2j is calculated to best fit our measurements at  $T = 75$  K (Fig. 2f) with a Knudsen number of  $\text{Kn} \equiv l_{\text{ee}}/W = 0.16$ . This is consistent with hydrodynamic electron flow in which  $l_{\text{ee}}$  is the smallest length scale in the system, in agreement with previous transport measurements<sup>3,6,11</sup>. The  $j_x$  and  $E_y$  profiles calculated for these parameters (Fig. 2j) are similarly curved (deviation scales as  $(l_{\text{ee}}/W)^{-2}$ , consistent with equation (1)), showing that the imaged  $E_y$  profile (Fig. 2f) approximates the actual Poiseuille  $j_x$  profile to within 5% (see the right  $y$  axis). The theoretical  $j_x$  profile corresponding to the  $T = 75$  K measurement does not reach zero at the walls. Extrapolating this profile to zero yields an estimated slip length<sup>19,32</sup> of  $l_{\text{slip}} \approx 500 \text{ nm}$ .

Having imaged the emergence of Poiseuille flow at increased temperatures, we now explore its carrier density dependence. For a linearly dispersing spectrum, Fermi liquid theory predicts  $l_{\text{ee}} \propto E_F/T^2 \propto \sqrt{|n|}/T^2$



**Fig. 3 | Carrier density dependence of hydrodynamic electron flow profiles.**

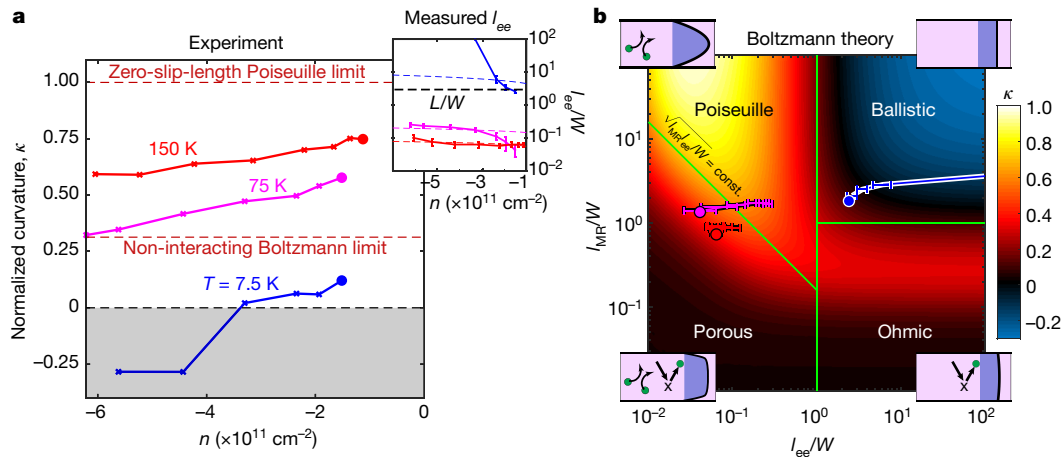
**a**,  $l_{MR}/W$  for  $T = 75$  K taken from Fig. 1d, with dots indicating values of  $n$  corresponding to experiments in subsequent panels. Between the green dots,  $l_{MR}$  is practically independent of  $n$  owing to the combination of phonon and impurity scattering. **b**, Comparison of magnetoresistance in units of the inverse transport mean free path  $W/l_{tr}$ , where for Dirac electrons  $l_{tr}(B) = h/[2e^2(\pi|n|)^{1/2}\rho_{xx}(B)]$  (where  $h$  is Planck's constant), at  $T = 75$  K for several

values of  $n$  indicated by the colour of the curve (corresponding to dots in **a**). The two green curves at higher  $|n|$  exhibit nearly indistinguishable magnetotransport. **c**,  $E_y/E_d$  profiles imaged for the same values of  $n$  as in **b** as indicated by colour ( $W/R = 1.3$  for each curve, with  $B = \pm 25.4$  mT, 18.5 mT and 12.5 mT for  $n = -6.2 \times 10^{11} \text{ cm}^{-2}$ ,  $-3.3 \times 10^{11} \text{ cm}^{-2}$  and  $-1.5 \times 10^{11} \text{ cm}^{-2}$ , respectively), demonstrating the monotonic increase of curvature with decreasing  $|n|$ .  $E_y^d$  is the bulk value for the classical Hall field,  $(B/ne)/W$ .

(where  $E_F$  is the Fermi energy), and so a variation of the flow profiles with  $n$  is expected. Varying  $n$ , however, will generically also change  $l_{MR}$ , possibly masking the relatively weak  $\sqrt{|n|}$ -dependence of  $l_{ee}$ . Fortunately, at elevated temperatures there is a range of  $n$  over which  $l_{MR}$  remains nearly constant owing to the compensating effects of phonon and impurity scattering (between green dots in Fig. 3a,  $T = 75$  K). In fact, the magnetoresistance at two substantially different densities (Fig. 3a, green dots) is nearly identical (Fig. 3b, green curves), implying that from transport measurements alone it is impossible to distinguish between electron flows at these densities (see Methods and Extended Data Fig. 6). However, the corresponding imaged  $E_y$  profiles (Fig. 3c, green curves)

are markedly different, varying in curvature by about 50%, which reflects the variation in  $l_{ee}$ . This result again highlights that  $E_y$  is a sensitive indicator for hydrodynamics. At even lower  $|n|$  (black dot, Fig. 3a)  $l_{MR}$  drops and both magnetoresistance (Fig. 3b, black curve) and the  $E_y$  profile (Fig. 3c, black curve) change as compared to higher densities.

We now systematically investigate how the curvature of  $E_y$  varies over a broader range of  $n$  and  $T$ . For each  $n$  and  $T$  we image the  $E_y$  profile, fit it to the form  $E_y(y) = ay^2 + c$  for  $|y/W| < 0.3$ , and extract the normalized curvature  $\kappa = -(a/c)/(W/2)^2$  ( $\kappa = 0$  for a flat profile and  $\kappa = 1$  for an ideal parabolic Poiseuille profile, reaching zero at the walls). Figure 4a plots the measured  $\kappa$  as a function of  $n$  for  $T = 7.5$  K,  $T = 75$  K and  $T = 150$  K.



**Fig. 4 | Curvature of the imaged  $E_y$  profiles and phase diagram of electron flow regimes.**

**a**, Normalized curvature,  $\kappa$ , of the imaged  $E_y$  profiles as a function of  $n$  and  $T$  as described in the main text (data points marked by crosses). Dashed red lines mark the maximal curvature obtained for non-interacting electrons based on Boltzmann calculations, and also the curvature of the ideal Poiseuille flow with zero slip length. Inset,  $l_{ee}$  at the values of  $n$  and  $T$  from the experiment (solid lines), determined by comparing the imaged  $E_y$  profiles to those calculated using the Boltzmann equations (error bars correspond to the standard deviation of  $l_{ee}$ , computed by least-squares fitting of Boltzmann calculations to experimental data). The coloured dashed lines are the corresponding predictions for  $l_{ee}$  based on many-body calculations for monolayer graphene<sup>33</sup>. The dashed black line marks the length of the device  $L$  (normalized by  $W$ ), above which the Boltzmann theory for an infinitely long channel can no longer predict  $l_{ee}$ . **b**, Phase diagram of electron flow as obtained from  $\kappa$ , calculated by Boltzmann theory (colour scale) as a function of  $l_{MR}/W$  and  $l_{ee}/W$ . The curvature values are determined by convolving the calculated profiles with the point spread function of the experiment at the same finite magnetic fields as in the experiment ( $W/R_c = 1.3$ ) for best comparison. The different electron flow regimes are labelled (ballistic, Ohmic, Poiseuille and

porous) together with illustrations of the relevant scattering mechanisms. Electrons are drawn as green circles, and  $E_y$  profiles are schematically drawn in purple. In the ballistic regime, the  $E_y$  profile is flat or even negatively curved (the magnitude of negative curvature is limited by the nonzero magnetic field). In the Ohmic regime, electrons scatter primarily from impurities/phonons (drawn as crosses), and the  $E_y$  profile can be gently curved. In the Poiseuille regime, electrons primarily scatter from other electrons, leading to a strongly parabolic  $E_y$  profile. In the porous regime, both electron scattering from impurities and phonons and electron–electron scattering have a prominent role, resulting in an  $E_y$  profile that is gently curved in the middle of the channel and reaches zero over a distance of the order of  $D_v = \frac{1}{2}\sqrt{l_{MR}l_{ee}}$  from the walls. The green lines mark the transitions between the different regimes: ballistic to Ohmic at  $l_{MR}/W = 1$ , transition to hydrodynamics  $l_{ee}/W = 1$ , and transition from Poiseuille to porous at  $D_v/W \approx 1$ . In the Poiseuille regime the profiles can reach a maximum curvature of  $\kappa = 1$ . The overlaid blue, purple and red paths correspond to the values of  $l_{MR}$  and  $l_{ee}$  (same error bars as in the inset of **a**) at  $T = 7.5$  K,  $T = 75$  K and  $T = 150$  K, respectively, from the experimental traces in **a**, with the dots indicating the lowest density.



At  $T = 7.5$  K we find that  $\kappa$  is close to zero, and even becomes negative at high density. We further observe that the value of  $\kappa$  monotonically increases with increasing  $T$  and decreasing  $|n|$ , with the measured curvature approaching the ideal Poiseuille value at the highest  $T$  and lowest  $|n|$ .

To demonstrate the relation between the curvature of the  $E_y$  profiles and the flow regime more quantitatively, we plot in Fig. 4b a phase diagram of the flow based on  $\kappa$  calculated using the Boltzmann theory as a function of the two length scales that control the physics:  $l_{\text{MR}}/W$  and  $l_{\text{ee}}/W$ . The phase space is demarcated into four regions: Ohmic, ballistic, Poiseuille and porous, the last two of which are hydrodynamic. In the Ohmic regime the curvature is small, peaking when  $l_{\text{MR}}/W \approx 0.25$ . In the ballistic regime, where  $l_{\text{MR}}/W > 1$ ,  $\kappa$  is governed by the reciprocal sum

$\left(\frac{1}{l_{\text{ee}}} + \frac{1}{l_{\text{MR}}}\right)^{-1}$ , and even becomes negative (see Fig. 4a at  $T = 7.5$  K). In the

left half of the phase diagram ( $l_{\text{ee}}/W < 1$ ), the flow is hydrodynamic, and is either Poiseuille (top left) or porous (bottom left) in character. The transition occurs when the so-called ‘Gurzhi’ length,  $D_v = \frac{1}{2}\sqrt{l_{\text{ee}}l_{\text{MR}}}$ , crosses  $W$ . In the porous regime ( $D_v < W$ ), named in analogy to water flow through porous media, both  $l_{\text{MR}}$  and  $l_{\text{ee}}$  can be smaller than  $W$ . Here,  $\kappa$  is low as in the Ohmic regime, but electron–electron interactions cause a sharp drop of  $E_y$  at the walls. In the Poiseuille regime ( $D_v > W$ ),  $\kappa$  increases substantially, approaching  $\kappa = 1$ , with the parabolic profiles of both  $E_y$  and  $j_x$  reaching zero at the walls (see Methods and Extended Data Fig. 8).

We now quantitatively compare the imaged  $E_y$  profiles at each density and temperature against the Boltzmann theory. Using the  $l_{\text{MR}}$  presented in Fig. 1d, we fit the entire Boltzmann profiles to our imaged profiles to determine the  $l_{\text{ee}}$  that gives the best match. The extracted values of  $l_{\text{ee}}$  (solid lines in the inset of Fig. 4a) are in close agreement with the many-body calculation for monolayer graphene<sup>33</sup> (see dashed lines in the inset of Fig. 4a), exhibiting the predicted decrease of  $l_{\text{ee}}$  with decreasing  $|n|$  and increasing  $T$ . Note that once  $l_{\text{ee}}$  exceeds the length of the channel (dashed black line) the Boltzmann calculations, which assume an infinite channel, lose their predictive power. Also, although at  $T = 7.5$  K and  $T = 75$  K the Boltzmann profiles closely match the overall magnitude and curvature of the imaged  $E_y$  profiles, at  $T = 150$  K, the best-fit profiles underestimate the imaged curvature (see Methods and Extended Data Fig. 9). This is probably caused by the scattering time approximation used in the calculation, suggesting that an improved microscopic understanding of electron–electron interactions is necessary to more completely understand hydrodynamics in real electronic systems (for example, using scattering integrals that better account for energy–momentum conservation in two dimensions, such as the long-lived odd-parity Fermi surface excitation modes proposed in refs. 34,35). Finally, we overlay the values of  $l_{\text{MR}}$  and  $l_{\text{ee}}$  obtained from the measurements onto Fig. 4b (coloured paths correspond to the different temperatures, dots indicate lowest densities), showing the trajectories through the phase diagram explored in the experiment. Probing deeper into the Ohmic regime is limited, as further decreasing  $l_{\text{MR}}$  requires low carrier densities where inhomogeneity near the channel edges becomes important ( $\Delta n_{\text{edges}} \approx 10^{10} \text{ cm}^{-2}$ ). Reaching deeper into the Poiseuille regime is also problematic, as the necessary higher temperature induces increased phonon scattering, resulting in  $D_v < W$ .

In conclusion, we have imaged electron flow through graphene channel devices by mapping the transverse component of the Hall electric field, which we find to be the essential element for distinguishing hydrodynamic from ballistic flow. With increasing temperature, we observe the evolution from flat ballistic profiles to curved profiles, producing images of Poiseuille electronic flow. Taken together with previous studies<sup>2–11</sup>, our experiments firmly establish the existence of an electron liquid that flows according to a universal hydrodynamic description. Our approach should enable further exploration of the physics of strongly interacting electrons upon application to other materials and topologically distinct flow geometries.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1788-9>.

- Gurzhi, R. N. Minimum of resistance in impurity free conductors. *Sov. Phys. JETP* **27**, 1019 (1968).
- de Jong, M. J. M. & Molenkamp, L. W. Hydrodynamic electron flow in high-mobility wires. *Phys. Rev. B* **51**, 13389–13402 (1995).
- Bandurin, D. A. et al. Negative local resistance due to viscous electron backflow in graphene. *Science* **351**, 1055–1058 (2016).
- Crossno, J. et al. Observation of the Dirac fluid and the breakdown of the Wiedemann–Franz law in graphene. *Science* **351**, 1058–1061 (2016).
- Moll, P. J. W., Kushwaha, P., Nandi, N., Schmidt, B. & Mackenzie, A. P. Evidence for hydrodynamic electron flow in PdCoO<sub>2</sub>. *Science* **351**, 1061–1064 (2016).
- Krishna Kumar, R. et al. Superballistic flow of viscous electron fluid through graphene constrictions. *Nat. Phys.* **13**, 1182–1185 (2017).
- Braem, B. A. et al. Scanning gate microscopy in a viscous electron fluid. *Phys. Rev. B* **98**, 241304 (2018).
- Gooth, J. et al. Thermal and electrical signatures of a hydrodynamic electron fluid in tungsten diphosphide. *Nat. Commun.* **9**, 4093 (2018).
- Gusev, G. M., Levin, A. D., Levinson, E. V. & Bakarov, A. K. Viscous transport and Hall viscosity in a two-dimensional electron system. *Phys. Rev. B* **98**, 161303 (2018).
- Bandurin, D. A. et al. Fluidity onset in graphene. *Nat. Commun.* **9**, 4533 (2018).
- Berdugin, A. I. et al. Measuring Hall viscosity of graphene’s electron fluid. *Science* **364**, 162–165 (2019).
- Elia, L. et al. Simultaneous imaging of voltage and current density of flowing electrons in two dimensions. *Nat. Nanotechnol.* **14**, 480–487 (2019).
- Lucas, A. & Fong, K. C. Hydrodynamics of electrons in graphene. *J. Phys. Condens. Matter* **30**, 053001 (2018).
- Levitov, L. & Falkovich, G. Electron viscosity, current vortices and negative nonlocal resistance in graphene. *Nat. Phys.* **12**, 672–676 (2016).
- Mohseni, K., Shakouri, A., Ram, R. J. & Abraham, M. C. Electron vortices in semiconductor devices. *Phys. Fluids* **17**, 100602 (2005).
- Andreev, A. V., Kivelson, S. A. & Spivak, B. Hydrodynamic description of transport in strongly correlated electron systems. *Phys. Rev. Lett.* **106**, 256804 (2011).
- Alekseev, P. S. Negative magnetoresistance in viscous flow of two-dimensional electrons. *Phys. Rev. Lett.* **117**, 166601 (2016).
- Falkovich, G. & Levitov, L. Linking spatial distributions of potential and current in viscous electronics. *Phys. Rev. Lett.* **119**, 066601 (2017).
- Torre, I., Tomadin, A., Geim, A. K. & Polini, M. Nonlocal transport and the hydrodynamic shear viscosity in graphene. *Phys. Rev. B* **92**, 165433 (2015).
- Pellegrino, F. M. D., Torre, I., Geim, A. K. & Polini, M. Electron hydrodynamics dilemma: whirlpools or no whirlpools. *Phys. Rev. B* **94**, 155414 (2016).
- Ho, D. Y. H., Yudhistira, I., Chakraborty, N. & Adam, S. Theoretical determination of hydrodynamic window in monolayer and bilayer graphene from scattering rates. *Phys. Rev. B* **97**, 121404 (2018).
- Levchenko, A., Xie, H.-Y. & Andreev, A. V. Viscous magnetoresistance of correlated electron liquids. *Phys. Rev. B* **95**, 121301 (2017).
- Lucas, A. & Hartnoll, S. A. Kinetic theory of transport for inhomogeneous electron fluids. *Phys. Rev. B* **97**, 045105 (2018).
- Scaffidi, T., Nandi, N., Schmidt, B., Mackenzie, A. P. & Moore, J. E. Hydrodynamic electron flow and Hall viscosity. *Phys. Rev. Lett.* **118**, 226601 (2017).
- Shytov, A., Kong, J. F., Falkovich, G. & Levitov, L. Particle collisions and negative nonlocal response of ballistic electrons. *Phys. Rev. Lett.* **121**, 176805 (2018).
- Alekseev, P. S. & Semina, M. A. Ballistic flow of two-dimensional interacting electrons. *Phys. Rev. B* **98**, 165412 (2018).
- Svintsov, D. Hydrodynamic-to-ballistic crossover in Dirac materials. *Phys. Rev. B* **97**, 121405 (2018).
- Narozhny, B. N., Gornyi, I. V., Mirlin, A. D. & Schmalian, J. Hydrodynamic approach to electronic transport in graphene. *Ann. Phys.* **529**, 1700043 (2017).
- Masubuchi, S. et al. Boundary scattering in ballistic graphene. *Phys. Rev. Lett.* **109**, 036601 (2012).
- Wang, L. et al. One-dimensional electrical contact to a two-dimensional material. *Science* **342**, 614–617 (2013).
- Holder, T. et al. Ballistic and hydrodynamic magnetotransport in narrow channels. Preprint at <https://arxiv.org/abs/1901.08546> (2019).
- Kiselev, E. I. & Schmalian, J. Boundary conditions of viscous electron flow. *Phys. Rev. B* **99**, 035430 (2019).
- Principi, A., Vignale, G., Carrega, M. & Polini, M. Bulk and shear viscosities of the two-dimensional electron liquid in a doped graphene sheet. *Phys. Rev. B* **93**, 125410 (2016).
- Gurzhi, R. N., Kalinenko, A. N. & Kopeliovich, A. I. Electron–electron collisions and a new hydrodynamic effect in two-dimensional electron gas. *Phys. Rev. Lett.* **74**, 3872–3875 (1995).
- Ledwith, P., Guo, H. & Levitov, L. The hierarchy of excitation lifetimes in two-dimensional Fermi gases. *Ann. Phys.* **411**, 167913 (2019).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

### Device fabrication

Scanning SET devices are fabricated using a nanoscale assembly technique<sup>36</sup>. The graphene/hBN devices are fabricated using electron-beam lithography and standard etching and nanofabrication procedures<sup>3</sup> to define the channels and evaporation of Pt (see main text) and Pd/Au (Extended Data Fig. 5) to deposit contact electrodes.

### Measurements

The measurements are performed on multiple graphene devices in two separate, home-built, variable-temperature, Attocube-based scanning probe microscopes. The microscopes operate in vacuum inside liquid helium dewars with superconducting magnets, and are mechanically stabilized using Newport laminar flow isolators. A local resistive surface mount device heater is used to heat the samples under study from  $T = 7.5$  K to  $T = 150$  K, and a DT-670-BR bare chip diode thermometer mounted proximally to the samples and on the same printed circuit boards is used for precise temperature control. The voltage imaging technique employed is presented in ref.<sup>12</sup>. Voltages and currents (for both the SET and sample under study) are sourced using a home-built digital-to-analog converter array, and measured using a home-built, software-based audio-frequency lock-in amplifier consisting of 1  $\mu$ V accurate d.c. and a.c. sources and a Femto DPLCA-200 current amplifier and NI-9239 analog-to-digital converter. The local gate voltage of the SET is dynamically adjusted via custom feedback electronics employing a least-squares regression algorithm to prevent disruption of the SET's working point during scanning and ensure reliable measurements.

The voltage excitations applied to the graphene channels are as follows: <4.3 mV at  $T = 7.5$  K, <7.5 mV at  $T = 75$  K, and <15 mV at  $T = 150$  K, all chosen to not cause additional current heating (Extended Data Fig. 4). The magnetic fields applied are in the range  $\pm 100$  mT.

### Determination of the momentum-relaxing mean free path

For a channel geometry of width  $W$ , as used in the experiments in this paper, the longitudinal resistivity,  $\rho_{xx}$ , reflects both the bulk resistivity of the graphene as well as scattering from the walls. To isolate the contribution from the bulk resistivity and determine the momentum-relaxing mean free path in the bulk,  $l_{MR}$ , we make use of the measured magnetoresistance. At any magnetic field we can obtain the transport mean free path from the measured  $\rho_{xx}$  via the Drude formula for Dirac electrons,  $l_{tr}(B) = h/[2e^2(\pi|n|)^{1/2}\rho_{xx}(B)]$ . In the semiclassical regime, the primary influence of a perpendicular magnetic field  $B$  is to bend the electron trajectories into cyclotron orbits of radius  $R_c = \frac{\hbar v_F}{eB}$ . At small magnetic fields such that the skipping orbit diameter is larger than the channel width,  $|W/R_c| < 2$ , electrons can be efficiently backscattered in the bulk and by the walls, and thus  $l_{tr}(B)$  contains the effects of both bulk and wall scattering. On the other hand, when  $|W/R_c| > 2$ , the backscattering from the walls is highly suppressed because a cyclotron orbit emerging from one wall cannot reach the other wall without scattering at least once in the bulk. In this regime the transport mean free path is primarily controlled by the bulk scattering length,  $l_{tr} \approx l_{MR}$ , with a small correction scaling as  $|W/R_c|^{-1}$  due to the volume participation ratio of skipping cyclotron orbits. In fact, using Boltzmann calculations of the magnetoresistance we can determine the correction factor over the entire phase space of the experiment. Extended Data Fig. 1 shows the ratio,  $l_{tr}/l_{MR}$ , calculated using Boltzmann theory (section 'Boltzmann simulations of flow profiles' in Methods), as a function of  $l_{ee}/W$  and  $l_{MR}/W$  for  $W/R_c = 3.2$ . By estimating the  $l_{ee}$  in our experiments using the  $E_y$  measurements and the Boltzmann calculations as in the main text (inset of Fig. 4a), and using  $l_{tr}$  as a zeroth-order estimate for  $l_{MR}$ , we can determine from Extended Data Fig. 1 the correction factor and obtain from our measured  $l_{tr}$  the bulk  $l_{MR}$ . Note that in the regions of the phase diagram traversed by the experiment (curves in Fig. 4b), the correction

factor is rather small and the maximum deviation of  $l_{MR}$  from  $l_{tr}$  is about 30%, so even the naive estimate,  $l_{MR} \approx l_{tr}$ , is already quite accurate.

### Diffusivity of etched channel walls in the experiment

Understanding the nature of electron scattering from the etched walls of the graphene channels is essential for both establishing the possibility of Poiseuille flow (diffusive walls are necessary for parabolic flow profiles), as well as for performing quantitative theoretical modelling of the imaging data to compare with experiment. In particular, we wish to know to what extent the scattering from the walls randomizes the momentum of an incoming electron. We quantify this property of the walls using a coefficient  $p$  that measures the probability of specular reflection, which can vary between zero and one. For perfectly specular walls,  $p = 1$ , and electrons will simply reflect off the walls in a mirror-like fashion. In the other limit, for perfectly diffusive walls,  $p = 0$ , and the momentum of the outgoing electron is completely randomized.

We use three different methods of increasing sophistication in order to extract the value of  $p$  for our channels, all of which indicate that the walls are strongly diffusive and  $p$  is nearly zero. To gain a basic intuition of the degree of diffusivity of the channel walls, we turn to a channel in which the walls for half of its length have been intentionally roughened through lithographic patterning (Extended Data Fig. 2a). This sample geometry allows us to directly compare how the voltage drops along the two different halves of the channel. We plot the voltage drop imaged along the centre of the channel in Extended Data Fig. 2b. Tellingly, we note that the voltage drops linearly across the region spanning both lithographic wall patterns (dashed red line in Extended Data Fig. 2a), with no discernible difference between the two halves. We can thus conclude that the walls of the section of the channel with the straight-line etch pattern are essentially equally as rough as the intentionally roughened section, suggesting that  $p \approx 0$ .

We next use the magnetoresistance data (see Fig. 1c) measured at  $T = 7.5$  K to estimate  $p$ . The double-peaked structure in the magnetoresistance is a telltale sign of ballistic transport, but is only present if the channel walls are diffusive, that is,  $p < 1$ . In short, the mechanism leading to the double peaks in a ballistic channel is the bending of electron trajectories by the field which forces them to scatter off the diffusive walls. At zero magnetic field, electrons traverse in straight lines, some of which have a shallow angle with respect to the walls, and in the absence of bulk scattering can have a long mean free path. The magnetic field bends these trajectories and forces electrons to hit the walls after a distance proportional to the cyclotron radius. For specular walls, this will not affect the resistivity. However, for diffusive walls this will cause extra back-scattering, leading to the double-peaked structure. This effect is pronounced when the bulk mean free path is long. However, for a short mean free path, bulk scattering will dominate and the double peak transforms into a single peak. At the transition from single- to double-peaked, the transport mean free path at zero field obeys the relation<sup>37</sup>  $l_{tr}(B=0) \approx W/(1-p)$ . Extended Data Fig. 2c plots the transport mean free path as a function of magnetic field and carrier density, determined from the measured  $\rho_{xx}$  data in Fig. 1c. We see that the double-peaked structure becomes a single, broad peak as  $l_{MR}$  decreases with decreasing density,  $|n|$ . Applying the above formula, we find that  $p \approx 0 \pm 0.1$  as  $l_{tr} \approx W = 4.7 \mu\text{m}$  at the transition from single- to double-peaked spectrum.

Finally, we can independently estimate  $p$  from the scaling of  $l_{tr}(B=0)$  as a function of density using the theoretical description for flow through a channel as a function of  $p$  developed by Molenkamp and de Jong<sup>2</sup>. We numerically solve for the fan diagram plotted in Extended Data Fig. 2d using the values of  $l_{MR}$  at  $T = 7.5$  K from experiment, which shows how  $l_{tr}(B=0)$  varies with carrier density  $n$ . The bold, red trace corresponds to  $p = 1$ , and is therefore identical to  $l_{MR}$  versus  $n$ , as expected for a channel with perfectly specular walls. As  $p$  is decreased from unity, the transport mean free path decreases and the curves level out, becoming rather flat at  $p = 0$ . The bold black trace in Extended Data Fig. 2d

corresponds to our experimentally measured  $l_{tr}(B=0)$  and closely matches, though very slightly undershoots the prediction for,  $p=0$ . Although the fit to the  $p=0$  theory is good, the slight mismatch suggests that, while the channel walls in the experiment are nearly fully diffusive, there may be an edge scattering mechanism at play not captured by the simple specular coefficient used in ref. <sup>2</sup>. Nevertheless, based on the variation between curves at different  $p$ , we can estimate that for our channel  $|p| < 0.1$ , consistent with the above analyses.

We also note that although the above analysis was performed for data taken at  $T = 7.5$  K, since we expect the diffusivity of the walls only to increase as the temperature is increased, the above estimates for  $p$  are then valid for all temperatures in our experiments. Further, while one might expect  $p$  to have some variation with carrier density owing to the varying strength of p-n junctions near the channel walls, our data strongly suggest that  $p$  remains close to zero for the entire range of hole carrier densities explored in the experiment, because any deviation from zero would increase the rate of change of  $l_{tr}(B=0)$  with  $n$ , which is inconsistent with Extended Data Fig. 2d. Thus, we conclude that the etched walls of the graphene channels are effectively fully diffusive throughout the entire phase diagram of our experiment.

### Dependence of Hall field profile curvature on magnetic field

Our method for mapping the Hall field,  $E_y$ , relies on the application of a small perpendicular magnetic field,  $B$ , to produce a Hall signal that is measurable by the scanning SET. We must then verify that this measurement is in the linear response regime with respect to  $B$ , namely that  $B$  is low enough not to alter the  $E_y$  profile. Specifically, we aim to prove that the curvature of the  $E_y$  profiles,  $\kappa$ , which is a main observable in this work, is not altered by  $B$ . In Extended Data Fig. 3a, we present the curvature  $\kappa$  imaged at a constant carrier density as a function of magnetic field at three temperatures,  $T = 4$  K,  $T = 75$  K and  $T = 150$  K. The curvature is extracted as described in the main text by a parabolic fit to  $E_y$  over the centre of the channel.

We note two distinct regimes of how  $\kappa$  depends on  $B$ : for  $W/R_c > 2$ ,  $\kappa$  has a strong field dependence, whereas for  $W/R_c < 2$ ,  $\kappa$  is constant at each temperature. In the higher-field regime for  $W/R_c > 2$ , closed cyclotron orbits can fit within the width of the channel. This leads to a rich evolution of  $E_y$  profiles that are no longer simply parabolic, and is the topic of a future work. In the lower-field regime for  $W/R_c < 2$ , we see that the measured curvature is constant to within our measurement noise down to the lowest fields measured ( $W/R_c \approx 1$ ). Imaging closer to  $B = 0$  is increasingly challenging, as the signal-to-noise ratio of the measured Hall voltage decreases linearly with decreasing field. Extended Data Fig. 3b shows similar traces ( $\kappa$  versus  $W/R_c$ ) calculated using Boltzmann equations for the values of  $l_{MR}/W$  and  $l_{ee}/W$  corresponding to the experiment. We find good correspondence between the Boltzmann simulations and the experiment. Most importantly, in the low field regime for  $W/R_c < 2$ , the simulations confirm that  $\kappa$  is independent of  $B$  as observed in the experiments, and extend this observation down to  $B = 0$ . Based on these results, the value of  $W/R_c = 1.3$  used for the  $E_y$  profile imaging in the experiments in the main text is justified.

Having justified experimentally and with Boltzmann simulations that the profiles are unperturbed in the low field regime  $W/R_c < 2$ , we also argue from analytic reasons why the flow profile is not expected to vary at low magnetic fields. In the hydrodynamic regime, the curvature is  $\kappa \approx \frac{W^2}{D_v^2 \sinh\left[\frac{W}{4D_v}\right]}$ , where  $D_v = \frac{1}{2} \sqrt{l_{MR} l_{ee}}$  is the Gurzhi length<sup>31</sup>. For low magnetic fields the correction to  $D_v$  has the form  $1 - \frac{2l_{eff}^2}{B^2}$ , where  $1/l_{eff} = 1/l_{MR} + 1/l_{ee}$ . This correction goes as  $B^2$ , and will be relevant only when  $R_c$  is of the order of  $l_{eff}$ , which we are far from at  $W/R_c = 1.3$  and the values of  $l_{ee}$  and  $l_{MR}$  that we achieve in the experiment in the hydrodynamic regime. We can therefore conclude that the curvature  $\kappa$  is not dependent on magnetic field for the parameters of our experiment.

### Dependence of Hall field profile curvature on voltage excitation

In order to drive current through the graphene channel devices, we apply an oscillating bias voltage of amplitude  $V_{ex}$  between the electrical contacts to the device. This excitation can in principle induce heating of the electrons above the temperature of the cryostat, and as a result cause an increase in curvature of the Hall field profiles. While this effect can be used<sup>2</sup> instead of substrate heating, we avoid this approach here owing to the additional spurious effects it may have on the curvature. We therefore choose an excitation amplitude at each temperature that is sufficiently low to minimally influence the curvature of the imaged profiles, but still high enough to enable a robust measurement.

Extended Data Fig. 4 shows the curvature of the field profiles versus excitation amplitude  $V_{ex}$  applied to the graphene device for two temperatures,  $T = 7.5$  K in the ballistic regime (blue trace) and  $T = 75$  K in the hydrodynamic, Poiseuille regime (purple trace). The curvature is extracted by a parabolic fit to the imaged  $E_y$  Hall profile imaged across the channel at a fixed density and magnetic field as described in the main text. In the Poiseuille regime ( $T = 75$  K, density  $n = -3.3 \times 10^{11} \text{ cm}^{-2}$ ,  $W/R_c = 1.3$ ), we see that the curvature ( $\kappa \approx 0.5$ ) is essentially independent of the excitation at least up to  $V_{ex} = 11$  mV, and therefore the excitation does not influence the physics of the electron flow. In the ballistic regime ( $T = 7.5$  K,  $n = -1.5 \times 10^{11} \text{ cm}^{-2}$ ,  $W/R_c = 1.3$ ), we see a clear increase in the curvature with increasing excitation due to electron heating. Still, for an excitation of  $V_{ex} = 4.3$  mV,  $\kappa$  is nearly zero and far below the Boltzmann limit marking the transition to hydrodynamic flow. We can thus safely choose such a low excitation and robustly image ballistic electron flow through the channel, although the specific value of  $\kappa$  may still be somewhat influenced by the excitation. In the experimental data presented in the main text, for  $T = 7.5$  K, the excitation across the graphene device is chosen such that  $V_{ex} < 4.3$  mV, for  $T = 75$  K,  $V_{ex} < 7.5$  mV, and for  $T = 150$  K,  $V_{ex} < 15$  mV.

### Comparison of Hall field profile curvature for different devices

We establish the consistency of our results across a set of graphene channel devices and scanning SET probes. The measurements in this work were carried out on two separate graphene device microchips, each imaged with a different scanning microscope and different SET. This allows us to compare between measurements and establish their lack of sensitivity to details specific to a particular graphene device or experimental setup. We denote the device used throughout the main text as device A. The additional device measured, which we denote as device B, is a channel with  $W = 5 \mu\text{m}$ , and  $L = 42 \mu\text{m}$ , allowing us additionally to rule out aspect-ratio-dependent effects (aspect ratio about 3 for device A versus about 8 for device B).

To most easily compare between devices, we examine the curvature of the Hall field profiles imaged at similar SET-graphene device separations. We focus on the magnetic field dependence of the curvature at several different temperatures and densities. The results are shown in Extended Data Fig. 5. We compare first between measurements taken at  $T = 7.5$  K and  $n = -1.5 \times 10^{11} \text{ cm}^{-2}$  in device A and  $T = 4$  K and  $n = -6 \times 10^{11} \text{ cm}^{-2}$  in device B. We then repeat the same comparison, now at  $T = 75$  K for both devices and  $n = -3.3 \times 10^{11} \text{ cm}^{-2}$  for device A and  $n = -1 \times 10^{12} \text{ cm}^{-2}$  for device B. The point spread function of the SET has a similar influence on both devices, and the same valid channel region is chosen for the extraction of the curvature ( $|W/R_c| < 0.3$ ).

In the low-temperature measurement, we observe a similar overall shape in the  $W/R_c < 2$  region. The low-field curvature in device A levels off at a slightly higher value than that in device B. The latter can be attributed to the different densities, since, as observed in Fig. 4a, at  $T = 7.5$  K the curvature exhibits strong density dependence. The curvatures imaged at elevated temperature closely match each other over the full range of magnetic fields, with small residual differences that are consistent with the density dependence in Fig. 4a. This indicates that

# Article

the hydrodynamic features observed in this work are not specific to the particular graphene sample or channel dimensions being measured.

## Distinguishing electron flow regime from transport

The temperature and width dependence of the resistivity of a channel relates to one of the earliest predictions in the field of electron hydrodynamics made by Gurzhi<sup>1</sup>. Specifically, he recognized the influence of wall scattering on the total current across the transition from the ballistic regime ( $l_{ee}, l_{MR} > W$ ) to the Poiseuille regime ( $l_{ee} < \frac{W^2}{l_{ee}} < l_{MR}$ ) by increasing  $T$  and thus decreasing the viscosity, while keeping momentum-relaxing collisions negligible. As the transition is made, a decrease in resistivity is expected, and in the Poiseuille regime, one is then expected to observe  $W^3$  scaling of the conductance. It is important to note, however, that for this to occur, we must maintain throughout the crossover the more stringent condition that  $l_{MR}$  is always much greater than  $W^2/l_{ee}$  and  $l_{ee}$ . In an experiment on monolayer graphene, and with channel widths that are amenable to measurement with the current generation of scanning SETs, it is impossible to reach deeply enough into the Poiseuille regime to meet this requirement, since the increase of temperature that leads to the decrease in  $l_{ee}$  also leads to increased momentum-relaxing collisions with phonons (that is, at higher temperature the condition  $l_{MR} > W^2/l_{ee}$  breaks down).

In this situation, the question of which observables are available for measurement becomes very important. It turns out that here the curvature in  $E_y$  plays a crucial role: When one performs a Boltzmann simulation<sup>24</sup>, it can be seen that the dependence of the resistivity for the values of  $l_{ee}/W$  and  $l_{MR}/W$  corresponding to our experiment is fairly weak, and much less informative than the dependence of the curvature of  $E_y$  for the same values of these parameters. This is shown explicitly in Extended Data Fig. 6 (taken from ref. <sup>2</sup>), which plots the dependence of the effective scattering length  $L_{eff} \propto 1/\rho_{xx}$  on  $l_{ee}/W$ , and by extension, on  $T$ , for different values of  $l_{MR}/W$ . The coloured ellipses correspond to the phase space regions reached in our experiments. While the resistivity variation over the experimentally relevant parameter range is weak, the curvature in  $E_y$  can vary substantially. This stems from the fact that the curvature of the flow profile is a geometric quantity, which directly relates to the length scales in the problem. Indeed, it is possible to maintain  $I = \int j_x dy = \text{constant}$  for a fixed applied voltage while changing the curvature of  $j_x$  from fully flat to fully parabolic.

It must be emphasized here that the difficulty in extracting the flow regime from the resistivity is not a simple case of being able to measure the latter to greater precision, which would naively allow us to extract meaningful information from even a small change in the resistivity. This would indeed be the case if all the quantities that make up the total resistivity, namely  $l_{imp}$ ,  $l_{ph}$  and  $l_{ee}$ , were to have a firmly understood functional dependence on the control parameters  $n$ ,  $T$  and  $W$ . However, this is not the case in graphene, and one can construct many models that would end up giving the same nearly flat form of  $\rho_{xx}$  versus  $T$ . Therefore, even a careful fitting of data to theory would not yield definitive information.

In this context, we mention that by using a judiciously selected sample geometry, such as in ref. <sup>10</sup>, a negative minimum in the vicinity resistance  $R_v$  of a bilayer graphene sample has been observed. This minimum, however, is related to a crossover in the quantity  $l_{ee}/x$ , where  $x$  is the distance from an injection contact to an adjacent probe contact, and the minimum can be attributed to a geometric effect which is absent in a channel geometry.

## Boltzmann simulations of flow profiles

To model electron flow through the graphene channels, we employ an approach based on the Boltzmann equation<sup>2,5,38</sup> that incorporates the effects of both electron-impurity and electron-phonon scattering as well as electron-electron interactions<sup>24</sup>:

$$\partial_t f + \mathbf{v} \cdot \nabla_r f + \frac{e}{m} (\mathbf{E} + \mathbf{v} \times \mathbf{B}) \cdot \nabla_v f = \left. \frac{\partial f}{\partial t} \right|_{\text{scatt}} \quad (2)$$

where the scattering integral

$$\left. \frac{\partial f(\mathbf{r}, \mathbf{v})}{\partial t} \right|_{\text{scatt}} = - \frac{f(\mathbf{r}, \mathbf{v}) - n(\mathbf{r})}{\tau} + \frac{2}{v_F^2 \tau_{ee}} \mathbf{v} \cdot \mathbf{j}(\mathbf{r}) \quad (3)$$

has two contributions: one from momentum-relaxing scattering, with a rate  $1/\tau_{MR}$ , and one from momentum-conserving, electron-electron scattering, with a rate  $1/\tau_{ee}$ . This equation describes the evolution of the semiclassical occupation number  $f(\mathbf{r}, \mathbf{v})$  for a wave packet of dynamical mass  $m$  at position  $\mathbf{r}$  and velocity  $\mathbf{v}$ , where  $n(\mathbf{r}) = \langle f \rangle_v$  is the local charge density,  $\mathbf{j}(\mathbf{r}) = \langle \mathbf{f} \mathbf{v} \rangle_v$  is the local current density,  $\langle \dots \rangle_v$  is the momentum average, and  $\frac{1}{\tau} = \frac{1}{\tau_{MR}} + \frac{1}{\tau_{ee}}$ . For the sake of simplicity, we consider the case of a circular Fermi surface with  $\mathbf{v} = v_F \hat{\mathbf{p}}(\theta)$ , where  $\hat{\mathbf{p}}$  is the radial unit vector at angle  $\theta$  and  $v_F$  is the Fermi velocity. Mean free paths are then simply defined as  $l_{MR(ee)} = v_F \tau_{MR(ee)}$ . The term proportional to  $1/\tau_{ee}$  is the simplest momentum-conserving scattering term that can be written, assuming that the electrons relax to a Fermi-Dirac distribution shifted by the drift velocity<sup>39,40</sup>. This form allows for different rates of momentum-relaxing and momentum-conserving scattering while still being amenable to computation. Although first used in the context of two-dimensional electron gases with parabolic bands, it has also been applied to graphene<sup>41</sup>.

The justification for the scattering integral in equation (3) is twofold. First, we note that the experiments are always firmly in the Fermi liquid regime ( $E_F \gg k_B T$ ), where the phase space for electron-hole scattering is negligible. This is illustrated in Extended Data Fig. 7, which plots the density and temperature of the experimental  $E_y$  profiles (red, blue and purple lines) presented in Fig. 4, along with the boundary  $E_F = k_B T$  (black curve). Above this boundary lies the ambipolar electron-hole/Dirac fluid regime in which both electrons and holes are present and scattering between them must be considered. Our experiments lie far below this boundary in the degenerate Fermi liquid regime, where only carriers of a single type are present and thus scattering is unipolar.

Second, beyond the fact that our experiments are in the Fermi liquid regime, the primary difference between equation (3) and a more graphene-specific scattering integral is that in equation (3) we have neglected the enhancement of collinear scattering due to the linear spectrum, which has a logarithmic dependence on the fine structure constant<sup>41</sup>. However, for graphene encapsulated in hBN, the fine structure constant is of the order of one, and thus the enhanced collinear scattering may be neglected. Moreover, by definition, collinear scattering mainly relaxes energy and only weakly relaxes the momentum direction. Since the latter plays the dominant role in how electrons flow through a channel, it is therefore safe to neglect this correction.

We assume a sample that is of infinite length along the  $x$  axis (which is the direction of current flow), and of finite width  $W$  along the  $y$  axis. The magnetic field is applied along the  $z$  direction. Diffuse scattering at the boundaries is imposed by the following boundary condition:

$$\begin{aligned} f\left(y = +\frac{W}{2}, \pi \leq \theta < 2\pi\right) &= f_{\text{boundary}} \\ f\left(y = -\frac{W}{2}, 0 \leq \theta < \pi\right) &= -f_{\text{boundary}} \end{aligned} \quad (4)$$

where  $f_{\text{boundary}}$  is a constant that is independent of  $\theta$  and that must be determined self-consistently. This ensures a uniform probability density for the angle at which an outgoing electron leaves a given wall, as required for completely diffuse scattering. Note that  $f_{\text{boundary}} = 0$  at zero magnetic field<sup>12</sup> but is non-zero in general. More generally, one could consider a finite degree of specularly for boundary scattering, by taking



$$f\left(y = +\frac{W}{2}, \pi \leq \theta < 2\pi\right) = p(\theta)f\left(y = +\frac{W}{2}, -\theta\right) + f_{\text{boundary}}$$

$$f\left(y = -\frac{W}{2}, 0 \leq \theta < \pi\right) = p(\theta)f\left(y = -\frac{W}{2}, -\theta\right) - f_{\text{boundary}} \quad (5)$$

where  $p(\theta) \in [0, 1]$  is the degree of specularity for electrons at incidence angle  $\theta$ . Although our calculations were limited to  $p = 0$  to match the experiment (see ‘Diffusivity of etched channel walls in the experiment’ in Methods), we expect that adding a small amount of specularity would only gradually wash out both ballistic and hydrodynamic effects.

Equation (2) is supplemented by Gauss’s law with a charge density given by  $en(\mathbf{x})$ . The resulting integrodifferential equation is solved numerically using the method of characteristics<sup>40</sup> to invert the differential part of the equation, and an iterative method to solve the integral part.

We emphasize that the above kinetic approach does not imply a no-slip boundary condition for the current. Instead, equation (5) merely imposes randomization of the incoming momentum under boundary scattering. This condition is well suited for doped graphene (‘Diffusivity of etched channel walls in the experiment’ in Methods and ref. <sup>32</sup>) and smoothly interpolates between effectively no-slip conditions in the hydrodynamic regime and a sizable slip length for ballistic flow. For this reason, the precise value of the specularity coefficient in the calculation does not qualitatively change the solution<sup>31</sup>.

We also note that, importantly, the determination of the flow profile by means of the Boltzmann distribution function ensures that full information of the kinematics is retained. This includes exceptional trajectories in the ballistic limit. We resolve not only the long-lived trajectories which travel almost tangentially to the boundary, but also the boundary skipping orbits which impact the walls many times<sup>31</sup>.

### Relation between $E_y$ and $j_x$ in the hydrodynamic regime

In the hydrodynamic regime for a channel of bulk resistivity  $\rho_{xx}^{\text{bulk}}$  with diffusive walls, the Hall field  $E_y(y)$  across the channel at weak magnetic field calculated using the Boltzmann kinetic equation approach<sup>13,42,43</sup> is given by:

$$E_y(y) = \rho_H j_x - \frac{E_x \frac{2l_{ee}}{R_c} \cosh\left(\frac{y}{D_v}\right)}{\cosh\left(\frac{W}{2D_v}\right)} \quad (6)$$

where  $\rho_H = B/ne$  is the Hall resistivity and  $E_x$  is the electric field along the channel. Additionally, we calculate the corresponding current density as:

$$j_x(y) = \frac{E_x}{\rho_{xx}^{\text{bulk}}} \left( 1 - \frac{\cosh\left(\frac{y}{D_v}\right)}{\cosh\left(\frac{W}{2D_v}\right)} \right) \quad (7)$$

We then note the following identity:

$$\partial_y^2 j_x = -\frac{E_x}{\rho_{xx}^{\text{bulk}}} \left( \frac{1}{D_v} \right)^2 \left( \frac{\cosh\left(\frac{y}{D_v}\right)}{\cosh\left(\frac{W}{2D_v}\right)} \right) \quad (8)$$

This allows us to substitute equation (8) into equation (6), and using

the relation  $\rho_H = \frac{\rho_{xx}^{\text{bulk}} l_{MR}}{R_c}$  we find:

$$E_y = \rho_H \left( j_x + \frac{1}{2} l_{ee}^2 \partial_y^2 j_x \right) \quad (9)$$

### Comparison of theoretical $E_y$ and $j_x$ curvature

For a long, ballistic channel, there are only two relevant parameters that determine the flow profile: the bulk mean free path normalized by the channel width  $l_{MR}/W$  and the specularity of the walls  $p$ . The case for specular walls is trivial, and the flow in the channel will be completely homogenous. The more interesting, experimentally relevant case is for diffusive walls with  $p = 0$ , where electrons flowing in the bulk are scattered with mean free path  $l_{MR}$  and electrons near the edge of the channel will encounter increased scattering by the diffusive walls. This physics alone, even without any electron–electron scattering, will produce a current density that varies from the bulk of the channel to the edges. For  $p = 0$ , the profile of  $j_x$  can then only be a function of the single parameter  $l_{MR}/W$ , where for  $l_{MR}/W \ll 1$  the flow is Ohmic and for  $l_{MR}/W > 1$  the flow is ballistic.

For a very wide channel, the current profile should be flat, as increasing  $W$  while keeping  $l_{MR}$  fixed leads to  $l_{MR}/W \ll 1$ , creating Ohmic flow. In this regime, information about the diffusive walls does not propagate substantially into the bulk, as the scattering length is much shorter than the channel width.

In the extreme ballistic regime for  $l_{MR}/W \rightarrow \infty$ ,  $j_x$  will also be flat, as an electron scattered at one wall will reach the other wall without scattering, effectively transmitting information about the diffusive walls uniformly throughout the channel. However, for non-infinite  $l_{MR}/W$ , this is no longer true, and  $j_x$  will necessarily have a curved profile. This is the experimentally relevant regime, as most published experiments on ballistic channels are done with  $l_{MR}/W$  not much greater than 1 (we reach  $l_{MR}/W \approx 5$ ). This is illustrated in Extended Data Fig. 8a, where we plot the curvature of  $j_x$  as a function of  $l_{MR}/W$ . The blue curve is based on the analytical formulas by de Jong and Molenkamp<sup>2</sup>, while the red curve is produced by a Monte Carlo electron billiards simulation. The curvature is extracted as in the main text: we fit a parabola of the form  $j_x(y) = ay^2 + c$  to the central 60% of the channel, with curvature  $\kappa = -(a/c)$   $(W/2)^2$ . We see that the curvature of  $j_x$  can be substantial even for very large  $l_{MR}/W$ , exemplifying the difficulty in distinguishing hydrodynamic electron flow from ballistic flow based on the current profile  $j_x$ .

We further compare the phase diagram defined by the theoretical estimate for the curvature  $\kappa$  of the  $E_y$  profiles presented in Fig. 4b with the phase diagram defined by the theoretical curvature of the  $j_x$  current density profiles. This allows us to present a more complete relation between  $E_y$  and  $j_x$  for  $W/R_c = 1.3$  for each flow regime as a function of  $l_{MR}/W$  and  $l_{ee}/W$ . The phase diagrams are presented side by side in Extended Data Fig. 8b, c. The  $j_x$  curvature phase diagram is constructed similarly to the  $E_y$  phase diagram, fitting a parabola to the centre of the  $j_x$  profiles calculated from the Boltzmann model after convolution with the point spread function of the SET. Examining first the right, non-hydrodynamic half of the phase diagram, we again note the large difference between the curvature in the ballistic regime of  $E_y$  and  $j_x$ . Whereas  $E_y$  can be negatively curved,  $j_x$  is always positively curved, with high curvature throughout the ballistic regime. The crossover between the ballistic regime and the Ohmic regime is evident in both phase diagrams, although the  $j_x$  curvature simply decreases from ballistic to Ohmic, while  $E_y$  goes through a local maximum near the crossover. In the hydrodynamic regime, both phase diagrams are similar, with the curvature matching exactly in both limits of strongly Poiseuille and strongly porous hydrodynamic electron flow. This highlights the restoration of a local relation between  $E_y$  and  $j_x$ , which leads to a convergence between these quantities in the hydrodynamic regime.

### Comparisons of imaged $E_y$ profiles to simulations

To determine the best match to theory, we fit the entirety of each imaged  $E_y(y)$  profile over the range of  $|y|/W < 0.3$  to the profiles obtained from the Boltzmann calculations. As  $l_{MR}$  is already determined independently from the magnetoresistance measurements, this procedure gives us the corresponding values of  $l_{ee}$ . The fit of each profile therefore

# Article

has to match not only the curvature, but also the overall height, which is related to the total conductivity. In Extended Data Fig. 9 we present three representative imaged  $E_y$  profiles along with the best-fit Boltzmann profiles (green curves). While the theory matches the imaged profiles at  $T = 7.5$  K (blue curve) and  $T = 75$  K (purple curve) in both curvature and height, the profile at  $T = 150$  K (red curve) is clearly more curved than the best-match Boltzmann profile. As stated in the main text, this mismatch may be due to the relaxation time approximation for electron–electron interactions used in the Boltzmann calculations. Further theoretical developments are necessary to more completely explain the hydrodynamic flow profiles at higher temperatures. Still, we emphasize that this mismatch between theory and experiment does not affect the main observation in the paper, which is the observation of Poiseuille electron flow and its distinction from ballistic flow.

## Data availability

The data that support the plots and other analysis in this work are available upon request.

## Code availability

Computer code for reproducing the Boltzmann simulations and computing the electron-electron scattering length is available upon request.

36. Weissman, J. et al. Realization of pristine and locally tunable one-dimensional electron systems in carbon nanotubes. *Nat. Nanotechnol.* **8**, 569–574 (2013).

37. Thornton, T. J., Roukes, M. L., Scherer, A. & Van de Gaag, B. P. Boundary scattering in quantum wires. *Phys. Rev. Lett.* **63**, 2128–2131 (1989).  
38. Ditlefsen, E. & Lothe, J. Theory of size effects in electrical conductivity. *Phil. Mag.* **14**, 759–773 (1966).  
39. Molenkamp, L. W. & de Jong, M. J. M. Electron-electron-scattering-induced size effects in a two-dimensional wire. *Phys. Rev. B* **49**, 5038 (1994).  
40. Courant, R. & Hilbert, D. *Methods of Mathematical Physics II: Partial Differential Equations* 62–131 (Wiley, 2008).  
41. Kashuba, O., Trauzettel, B. & Molenkamp, L. W. Relativistic Gurzhi effect in channels of Dirac materials. *Phys. Rev. B* **97**, 2015129 (2018).  
42. Gurzhi, R. N. Hydrodynamic effects in solids at low temperature. *Sov. Phys. Usp.* **11**, 255–270 (1968).  
43. Alekseev, P. S. et al. Nonmonotonic magnetoresistance of a two-dimensional viscous electron-hole fluid in a confined geometry. *Phys. Rev. B* **97**, 085109 (2018).

**Acknowledgements** We thank G. Falkovich, A. Shytov, L. Levitov, D. Bandurin and R. Krishna-Kumar for discussions. We further acknowledge support from the Helmsley Charitable Trust grant, the ISF (grant number 712539), WIS-UK collaboration grant, the ERC-Cog (See-1D-Qmatter number 647413), the Deloro Award, the Minerva Foundation, and the Emergent Phenomena in Quantum Systems initiative of the Gordon and Betty Moore Foundation.

**Author contributions** J.A.S., L.E. and S.I. conceived the experiments. J.A.S., L.E., A.R., D.D. and S.I. performed the experiments. J.A.S., L.E., A.R. and S.I. analysed the data. J.B., D.J.P. and M.B.-S. fabricated the graphene devices. K.W. and T.T. supplied the hBN crystals. T.S., T.H., R.Q., A.P., A.R. and A.S. performed theoretical calculations. J.A.S., L.E. and S.I. wrote the manuscript, with input from all authors.

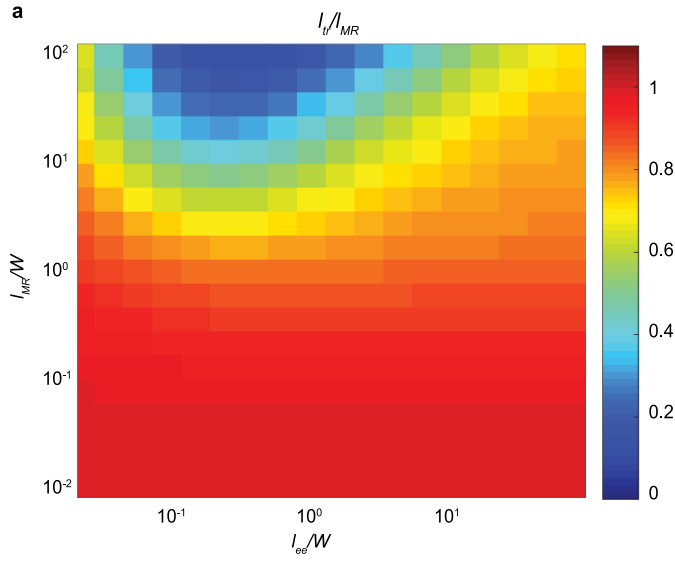
**Competing interests** The authors declare no competing interests.

## Additional information

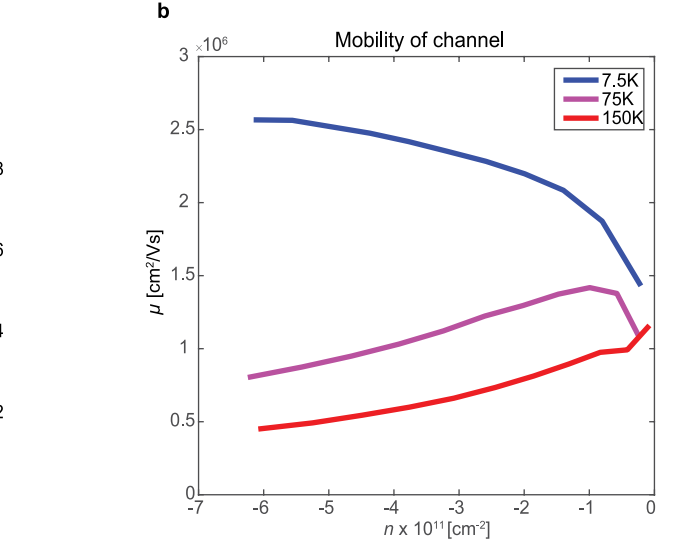
**Correspondence and requests for materials** should be addressed to S.I.

**Peer review information** *Nature* thanks Klaus Ensslin, Boris Narozhny and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

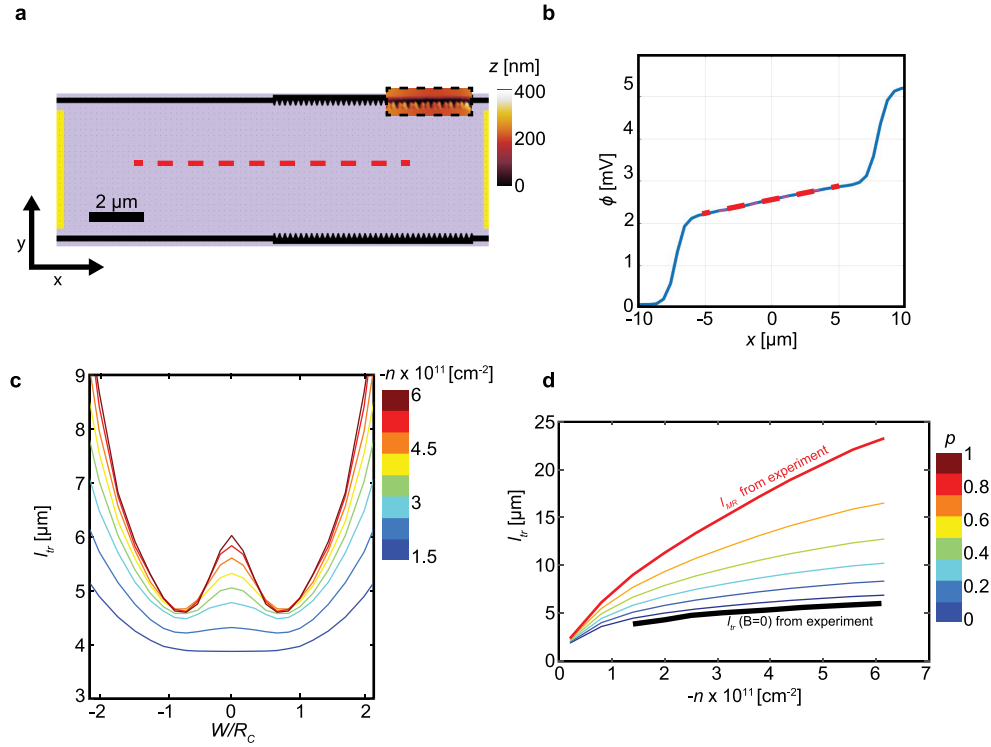
**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



**Extended Data Fig. 1 | Relation between transport and momentum-relaxing mean free path across the phase diagram of flow regimes and the electron mobility. a**, Boltzmann calculation of  $l_{tr}$  versus  $l_{MR}$  and  $l_{ee}$ . The two-dimensional map shows the ratio of the finite-field-transport mean free path,



$l_{tr}(B) = h/[2e^2 (\pi|n|)^{1/2} \rho_{xx}(B)]$ , and the bulk mean free path,  $l_{tr}/l_{MR}$ , calculated using Boltzmann theory at  $W/R_c = 3.2$  for a channel with diffusive walls, as a function of  $l_{ee}/W$  and  $l_{MR}/W$ . **b**, Electron mobility  $\mu$  measured with scanning SET, equivalent to the  $l_{MR}$  data presented in Fig. 1d.

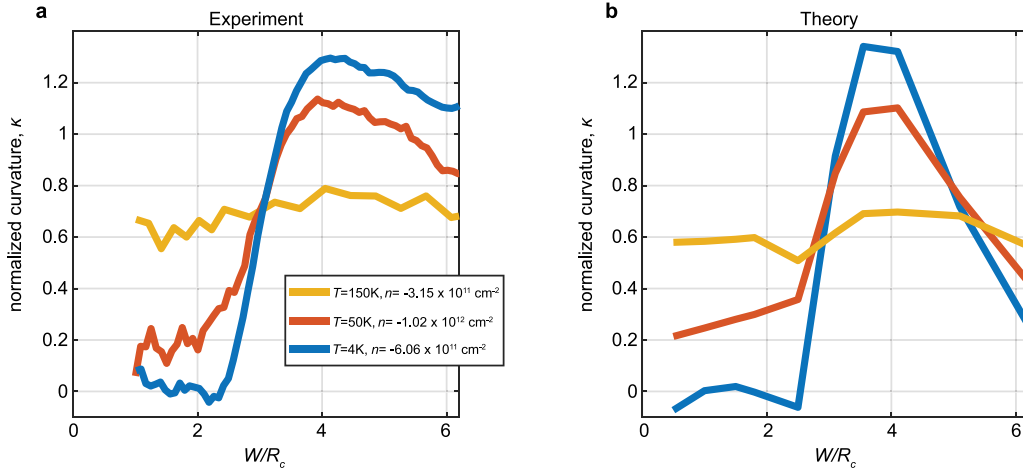


**Extended Data Fig. 2 | Diffusivity of etched channel walls in the experiment.**

**a**, Illustration of a channel used to assess the diffusivity of etched walls by direct comparison to lithographically roughened walls. The walls of the left half of the channel are patterned with a typical straight-line pattern, whereas the right half is patterned with a saw-toothed pattern to introduce roughness. The region enclosed by the dashed box in the upper right is an AFM image of the etched walls. The red dashed line marks the spatial region spanning both wall patterns along which the potential drop is measured. **b**, Measured potential

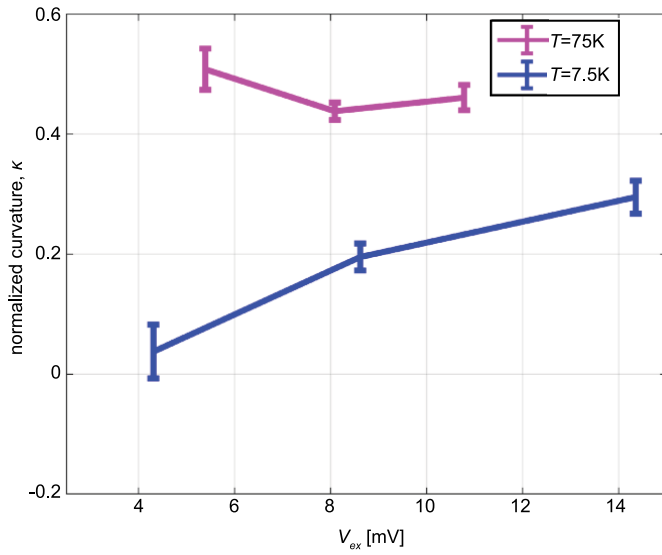
drop along the centre of the channel (red dashed line in **a**). Away from the voltage steps at the contacts, the potential drops linearly across the device, with no observable change in slope when the walls transition from straight line to saw-tooth etches. **c**, Zoom of magnetoresistance data from Fig. 1c plotted as  $l_{tr}(B)$  for varying density. At the transition where the double peak disappears,  $l_{tr}(B=0) \approx W/(1-p)$ , allowing estimation of specularity  $p$ . **d**, Theoretical scaling of  $l_{tr}(B=0)$  with  $n$  for varying  $p$  superimposed with the experimental data (bold black line), indicating that  $p$  is nearly zero (fully diffusive walls).



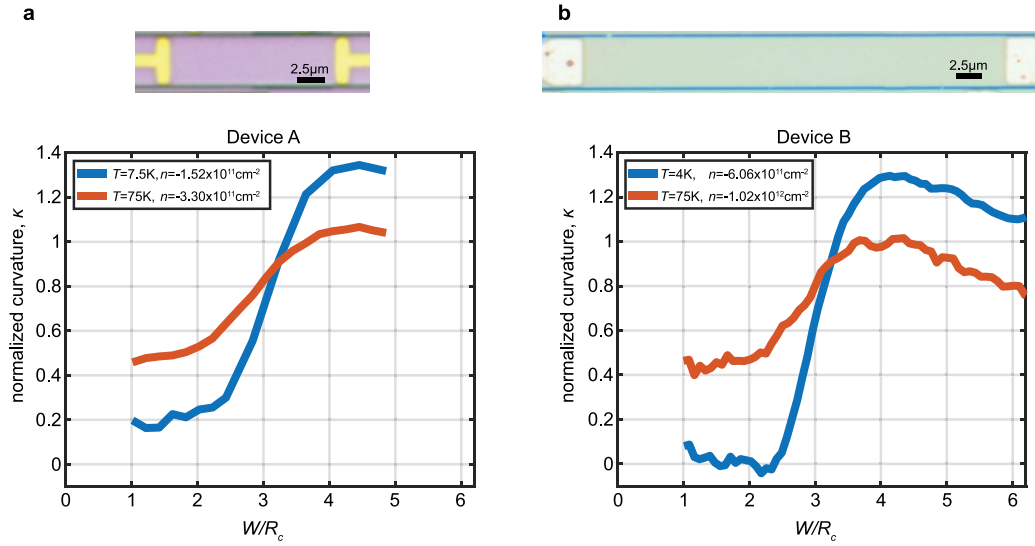


**Extended Data Fig. 3 | Dependence of Hall field profile curvature on magnetic field. a,** Measured traces of  $\kappa$ , extracted from  $E_y$  using a fit to the centre of the channel, as a function of magnetic field plotted in units of  $W/R_c \propto B$ . The blue trace is measured at  $T = 4 \text{ K}$  and hole density of  $n = -6.06 \times 10^{11} \text{ cm}^{-2}$  on device B (see Methods and Extended Data Fig. 5). The orange trace is measured at  $T = 50 \text{ K}$  and  $n = -1.02 \times 10^{12} \text{ cm}^{-2}$  on device B, and the yellow trace is measured at  $T = 150 \text{ K}$  and a hole density of  $n = -3.15 \times 10^{11} \text{ cm}^{-2}$  on

device A, which is the device used throughout the main text. Two distinct regimes are apparent: Below  $W/R_c \approx 2$ , the curvature is nearly independent of  $W/R_c$ , whereas above it varies noticeably, acquiring large values. **b,** Curvature as a function of  $W/R_c$  extracted from a Boltzmann simulation of  $E_y$  as described in the main text. Coloured curves correspond to values of  $I_{\text{MR}}$  and  $I_{\text{ee}}$  that best match experiment. This figure verifies that by imaging at  $W/R_c = 1.3$  as in the main text, the profiles are not influenced by the magnetic field.

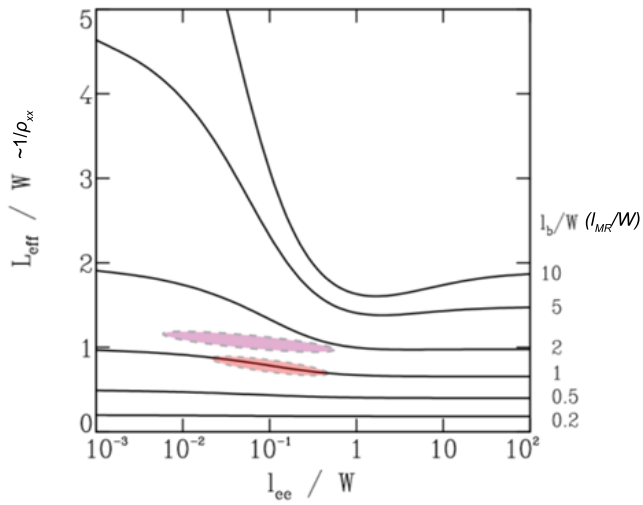


**Extended Data Fig. 4 | Dependence of Hall field profile curvature on voltage excitation.**  $\kappa$ , the normalized curvature of  $E_y$ , is plotted as a function of the excitation amplitude  $V_{ex}$  applied between the contacts of the channel. Error bars correspond to standard deviation of  $\kappa$  from the least-squares fit of a parabola to the data. The blue trace shows  $T=7.5$  K and  $n=-1.5 \times 10^{11} \text{ cm}^{-2}$ ; the purple trace shows  $T=75$  K and  $n=-3.3 \times 10^{11} \text{ cm}^{-2}$ . This plot verifies that by choosing appropriate values for the excitation, as was done for the experiments in the main text, electron heating effects are negligible.



**Extended Data Fig. 5 | Comparison of Hall field profile curvature for different devices.** **a**, Top, optical image of graphene device (device A) patterned into the geometry of a channel, with  $W = 4.7 \mu\text{m}$  and  $L = 15 \mu\text{m}$ , studied in the main text. Bottom, normalized curvature of  $E_y$ ,  $\kappa$ , measured as a function of  $W/R_c$ . **b**, Top, optical image of an additional graphene device

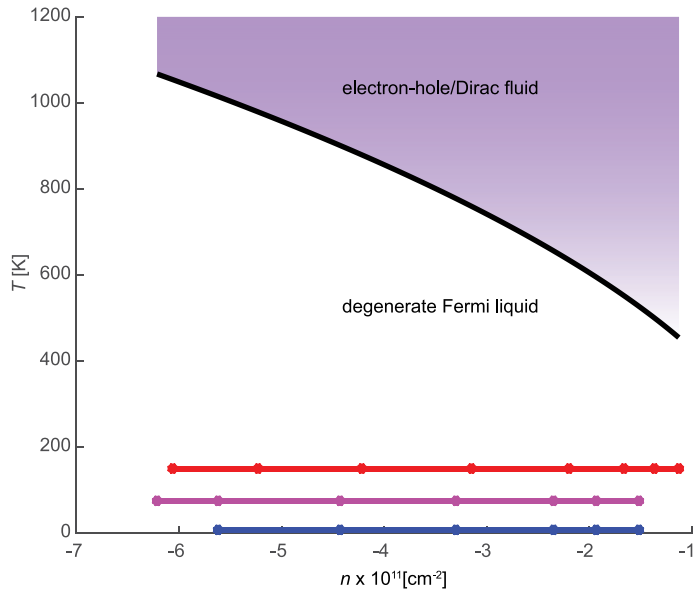
(device B) used for similar measurements, with  $W = 5 \mu\text{m}$  and  $L = 42 \mu\text{m}$ . This device was measured in a separate cryostat with a different scanning microscope and different SET. Colour differences between optical images are due to lighting conditions. Bottom,  $\kappa$  versus  $W/R_c$  measured for device B, showing a result highly consistent with that in **a**.



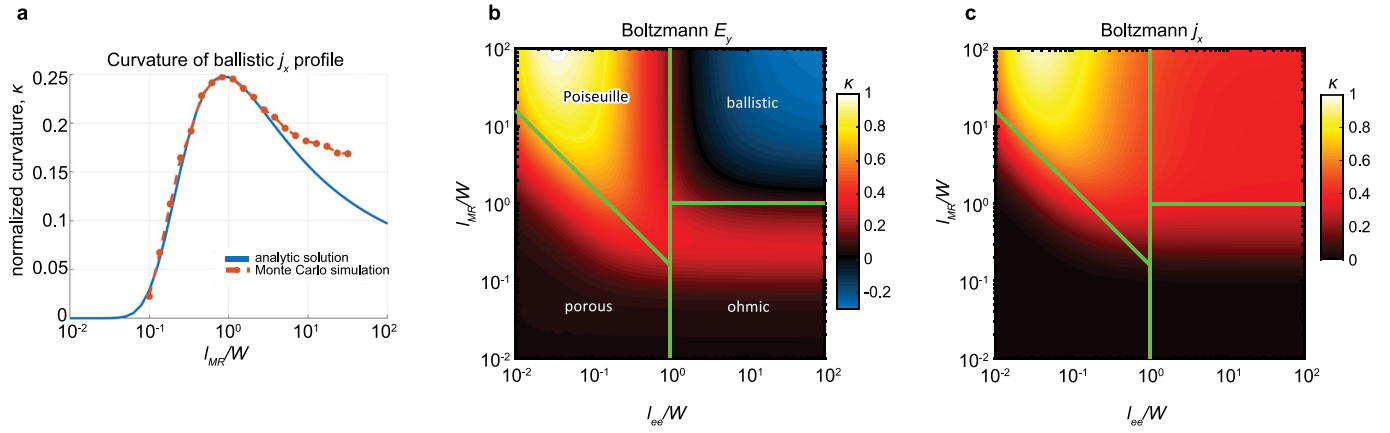
**Extended Data Fig. 6 | Distinguishing electron flow regime from transport.**

The graph shows the dependence of  $L_{eff}/W$ , which is inversely proportional to the resistivity, on  $l_{ee}/W$ , for fixed values of  $l_b/W = l_{MR}/W$ . The purple- and red-coloured regions correspond to the parameter ranges of our experiment for  $T = 75$  K and  $T = 150$  K, respectively. It is evident from these curves that the dependence of the resistivity on  $l_{ee}/W$  is fairly weak when  $l_{MR}/W$  is not much larger than 1. Figure reprinted with permission from de Jong and Molenkamp<sup>2</sup>; copyright 2019 by the American Physical Society.



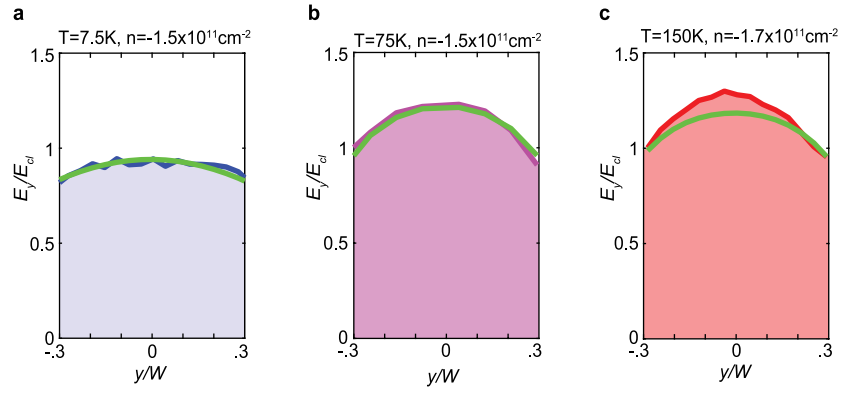


**Extended Data Fig. 7 | Phase diagram demonstrating that the experiment falls within the Fermi liquid regime.** The black line shows the equality  $E_F = k_B T$ , which separates the temperature–density plane into two distinct regimes. Above this line is the Dirac fluid regime in which electrons and holes are both present and thus electron–hole scattering must be considered. Below this line is the degenerate Fermi liquid regime in which only one charge carrier is present. The blue, purple and red lines correspond to the experiments presented in Fig. 4 at  $T = 7.5$  K,  $T = 75$  K and  $T = 150$  K, respectively, and show that our experiments are categorically within the Fermi liquid regime.



**Extended Data Fig. 8 | Curvature of ballistic  $j_x$  and comparison of theoretical  $E_y$  and  $j_x$  across phase diagram.** **a**, Curvature of ballistic current profile versus  $I_{MR}/W$ . The analytic solution (blue curve) is based on de Jong and Molenkamp<sup>2</sup> and the red curve is a Monte Carlo billiard ball simulation result. The two methods agree perfectly until  $I_{MR}$  exceeds the channel length used in the billiard ball simulation, beyond which the solutions begin to deviate. **b**, Curvature  $\kappa$  of  $E_y$ , as in Fig. 4b, calculated by Boltzmann simulation (see Methods), as a function of  $I_{ee}/W$  and  $I_{MR}/W$  for  $W/R_c = 1.3$ . Curvature is calculated over the centre of the channel. Green lines divide the panel into flow

regimes as in Fig. 4b. **c**, Curvature  $\kappa$  of  $j_x$ , extracted from the same simulation as **a**. For  $j_x$ , the curvature in the ballistic regime is essentially constant at  $\kappa \approx 0.31$  and so the curvature of  $j_x$  is less discriminating between the hydrodynamic and ballistic regimes than the curvature of  $E_y$ , which becomes negative. In the other regimes, the curvatures of  $j_x$  and  $E_y$  are very similar, and the differences between them diminish as each of the length scales becomes much smaller than  $W$ . In the hydrodynamic regime the curvature saturates on the maximal possible value for a strictly parabolic profile, and in the porous regime it follows the length scale  $D_v = \frac{1}{2} \sqrt{I_{MR} l_{ee}}$  as expected.



**Extended Data Fig. 9 | Comparisons of representative imaged  $E_y$  profiles to the Boltzmann simulated profiles.** Boltzmann simulation profiles are plotted in green, whereas the experimental data is plotted with the same colour

scheme as in main text based on temperature (blue for  $T=7.5\text{ K}$ , purple for  $T=75\text{ K}$  and red for  $T=150\text{ K}$ ). The field  $E_y$  is normalized as in the main text by the classical Hall field  $E_{cl} = (B/ne)(I/W)$ .

# Interlayer exciton laser of extended spatial coherence in atomically thin heterostructures

<https://doi.org/10.1038/s41586-019-1779-x>

Received: 31 December 2018

Accepted: 28 August 2019

Published online: 25 November 2019

Eunice Y. Paik<sup>1,4</sup>, Long Zhang<sup>1,4</sup>, G. William Burg<sup>2</sup>, Rahul Gogna<sup>3</sup>, Emanuel Tutuc<sup>2</sup> & Hui Deng<sup>1\*</sup>

Two-dimensional semiconductors have emerged as a new class of materials for nanophotonics owing to their strong exciton–photon interaction<sup>1,2</sup> and their ability to be engineered and integrated into devices<sup>3</sup>. Here we take advantage of these properties to engineer an efficient lasing medium based on direct-bandgap interlayer excitons in rotationally aligned atomically thin heterostructures<sup>4</sup>. Lasing is measured from a transition-metal dichalcogenide heterobilayer (WSe<sub>2</sub>–MoSe<sub>2</sub>) integrated in a silicon nitride grating resonator. An abrupt increase in the spatial coherence of the emission is observed across the lasing threshold. The work establishes interlayer excitons in two-dimensional heterostructures as a gain medium with spatially coherent lasing emission and potential for heterogeneous integration. With electrically tunable exciton–photon interaction strengths<sup>5</sup> and long-range dipolar interactions, these interlayer excitons are promising for application as low-power, ultrafast lasers and modulators and for the study of many-body quantum phenomena<sup>6</sup>.

Semiconductor lasers are ubiquitous in today's technology because they are compact, cover a wide range of wavelengths and allow efficient electrical pumping and fast electrical modulation. They are predominantly based on traditional III–V quantum wells. To achieve lower power consumption, more compact size and a higher degree of integration with silicon, there has been tremendous effort to develop alternative gain materials and structures, such as nanowire lasers<sup>7</sup>, spasers<sup>8</sup> and photonic crystal lasers<sup>9</sup>. However, tunability, electrical pumping and heterogeneous integration remain as common challenges.

Recently, monolayer transition-metal dichalcogenide crystals (TMDCs) have emerged as a new class of material for semiconductor lasers, as they are atomically thin and feature strong exciton emission<sup>1,2</sup>. Whereas lattice mismatch limits the choice of substrates for three-dimensional (3D) semiconductors, two-dimensional (2D) TMDCs do not have dangling bonds, and can be directly integrated with different substrates<sup>3</sup>. Previous studies have used two criteria to assess lasing in monolayer TMDCs: nonlinear intensity dependence, and linewidth reduction as a function of pump power<sup>10–15</sup>. However, the photon flux appears to be below the stimulated emission threshold<sup>16</sup>. Spatial coherence—an important property for characterizing lasers—has not been studied. Hence it is difficult to exclude localized excitons, such as point defects, as the source of the observed nonlinear power dependence. Moreover, with only a monolayer as the gain medium, tunability is limited and vertical p–n junctions are not possible without contacting with other doped semiconductors.

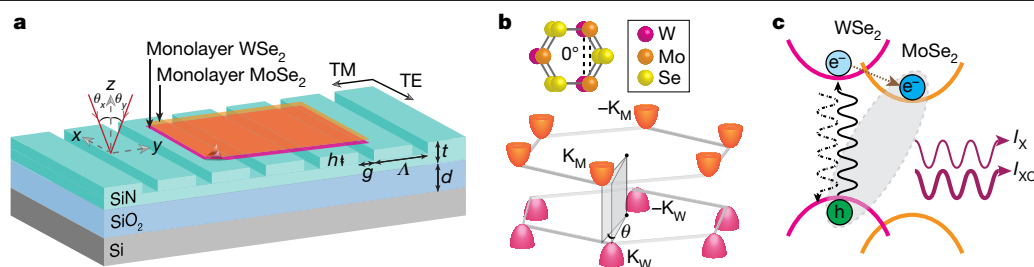
In contrast, heterostructures open the door to the engineering of band structures and exciton states. Spatially indirect excitons in heterostructures have been intensively studied<sup>5,17</sup>, for they feature

an electrically tunable static dipole with long-range dipole interactions, promising rich many-body quantum phenomena<sup>6</sup>. However, the reduced oscillator strength of spatially indirect excitons typically renders them dark and hard to access.

Here we show that in rotationally aligned 2D WSe<sub>2</sub>–MoSe<sub>2</sub> heterobilayers integrated on a silicon nitride (SiN) cavity (Fig. 1), interlayer excitons form an efficient gain medium, supporting lasing with extended spatial coherence at a low population inversion density. As illustrated in Fig. 1b, by forming a direct bandgap between the two monolayers<sup>4</sup> that are less than one nanometre apart, the interlayer excitons retain a sufficiently large oscillator strength. With type-II band alignment, the heterobilayer forms a three-level system that allows efficient pumping through the intralayer exciton resonances followed by rapid electron transfer to a lower-energy empty conduction band<sup>18,19</sup> (Fig. 1c). As a result, population inversion is readily achieved at the reduced bandgap while avoiding fast intralayer radiative loss of the carriers. Moreover, unlike some of the cavities used for monolayer exciton lasers, the cavity mode in our device fully covers the heterobilayer, allowing gain over the full area of the bilayer, and supporting extended spatial coherence (Fig. 1a). We observe lasing accompanied by an abrupt increase in the spatial coherence length as the photon occupancy exceeds unity. The emission intensity increases nonlinearly more than 100-fold across the threshold, and then continues to increase linearly with pump power (without saturation) up to the highest power used. Our results establish interlayer excitons in engineered TMDC heterobilayers as an efficient lasing medium, which, compared to excitons in monolayer TMDCs, feature electrically tunable long-range dipole interaction and oscillator strength<sup>5</sup>, robust valley polarization<sup>19</sup>, and a type-II band alignment

<sup>1</sup>Department of Physics, University of Michigan, Ann Arbor, MI, USA. <sup>2</sup>Microelectronics Research Center, Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, USA. <sup>3</sup>Applied Physics Program, University of Michigan, Ann Arbor, MI, USA. <sup>4</sup>These authors contributed equally: Eunice Y. Paik, Long Zhang. \*e-mail: dengh@umich.edu





**Fig. 1 | Illustration of the heterobilayer/grating-cavity laser system.**

**a**, Schematic of the laser device, consisting of a heterobilayer on a grating cavity. The along-bar (cross-bar) direction and polarization are defined as  $x$  ( $y$ ) and TE (TM) respectively. Grating cavity design parameters are the following: total SiN thickness ( $t$ ), SiO<sub>2</sub> thickness ( $d$ ), grating thickness ( $h$ ), grating period ( $\Lambda$ ) and gap width ( $g$ ). We define  $\theta_x$  ( $\theta_y$ ) as the azimuthal angle of the light beam in the  $x$ - $z$  ( $y$ - $z$ ) plane with respect to the  $z$  axis, as indicated by the red arrows. **b**, Illustration of the rotationally aligned heterobilayer with twist angle  $\theta = 0^\circ$  (top), and correspondingly a direct bandgap at the K valleys (bottom).

$K_M$  and  $K_W$  denote the K valleys of the MoSe<sub>2</sub> and WSe<sub>2</sub> layers, respectively. **c**, Band alignment and carrier dynamics of the heterobilayer. The heterobilayer has a type-II band alignment, forming a three-level system for the injected carriers. Intralayer excitons are excited by a pump laser in the WSe<sub>2</sub> layer (solid wavy line). Some electrons transfer to the lower MoSe<sub>2</sub> conduction band on a fast (10–100 fs) timescale (dotted line), while others recombine as intralayer excitons with lifetimes of 1–10 ps (dash-dotted wavy line). Without the cavity, the interlayer excitons (dashed line) recombine with a lifetime of the order of 1 ns ( $I_x$ ), and, with cavity enhancement, of the order of 100 ps ( $I_{xc}$ ).

well-suited for electrical injection via an atomically thin bilayer p–n junction<sup>20,21</sup>.

The lasing device comprises a rotationally aligned WSe<sub>2</sub>–MoSe<sub>2</sub> heterobilayer placed on a SiN grating resonator, as illustrated in Fig. 1a. To form bright interlayer excitons, we accurately align the crystal axes of the WSe<sub>2</sub> and MoSe<sub>2</sub> monolayers to within 1° of relative rotation, as verified by second-harmonic generation (SHG) measurements (Extended Data Fig. 1). Consequently, the band extrema at the K valleys of the two monolayers align in momentum space to form a direct bandgap (Fig. 1b). With type-II band alignment, carriers can be injected into the heterobilayers efficiently via the intralayer exciton resonance, followed by rapid electron transfer to the empty conduction band of MoSe<sub>2</sub> on a timescale of 10–100 fs (Fig. 1c)<sup>18,19</sup>. As a result, band inversion can be established at the smaller, interlayer bandgap. Once separated into the two monolayers, radiative recombination is reduced, rendering long interlayer exciton lifetimes of the order of 1 ns (Extended Data Fig. 2). Photoluminescence (PL) measurements of the heterobilayer show that interlayer exciton emission is much stronger than intralayer emission (Fig. 2b), confirming efficient charge transfer and sufficient build-up of the interlayer-exciton population. Spatially resolved PL shows uniform emission from the interlayer (intralayer) excitons in the bilayer (monolayer) regions (Extended Data Fig. 3).

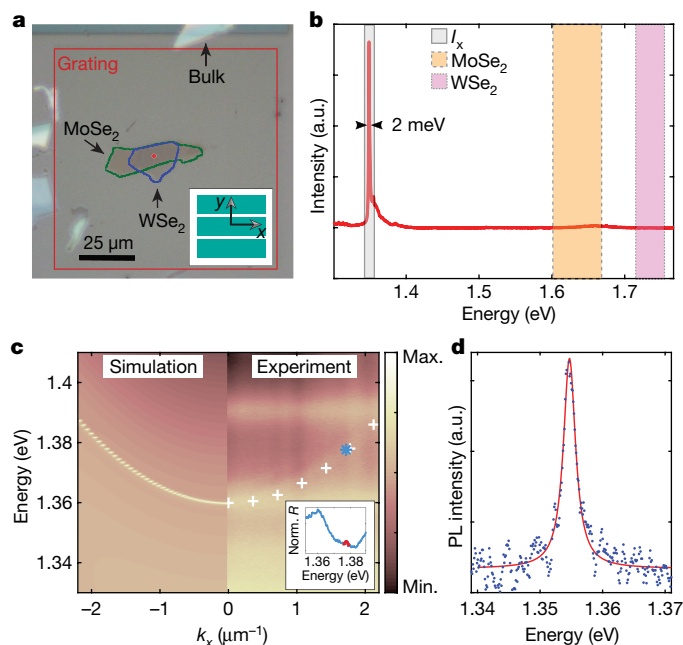
The grating cavity provides optical feedback when photons are coupled to its resonances. The cavity modes are sensitive to the propagation and the polarization directions of the electric field. We define the propagation (polarization) direction along the grating bar as  $x$  (TE) and across the bar as  $y$  (TM), as illustrated in Fig. 1a. We tune the grating period  $\Lambda$ , thickness  $h$  and fill factor  $g$  to obtain a high quality factor ( $Q$ -factor) for the TE mode and match it to the exciton resonance at zero in-plane wavenumber,  $k = 0$ . The heterobilayer lies directly on the grating where the evanescent field remains strong<sup>22</sup>. The TM cavity modes are far blue-detuned from the excitons; therefore, the TM exciton modes are not affected by the cavity (Extended Data Fig. 4).

We confirm the TE-cavity modes by measuring the empty-cavity dispersion with angle-resolved reflectance spectroscopy; the results agree well with the simulation by rigorous coupled wave analysis (RCWA), as shown in Fig. 2c. The TE mode  $Q$ -factor from the simulation is around 2,000. However, the actual cavity  $Q$ -factor is presumably lower, owing to fabrication imperfections. From the reflectance spectra of the empty cavity, we estimate a  $Q$ -factor of between 500 and 680 (Fig. 2c inset), but the exact value is difficult to determine owing to low contrast and white-light noise. The PL spectral linewidth from the device corresponds to a  $Q$ -factor of around 630 (Fig. 2d).

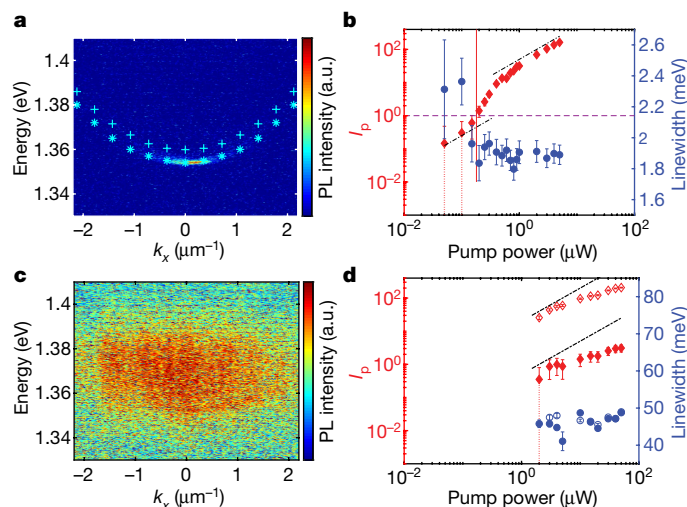
The heterobilayer allows efficient optical pumping through the intralayer exciton resonances, which are far above the resonances of the

interlayer excitons or the cavity. With the pump laser at 1.7 eV, the PL from the cavity mode at  $k \approx 0$  and energy  $E \approx 1.35$  eV brightens rapidly as the pump power increases, as seen in the along-bar angle-resolved PL (Fig. 3a).

Integrating over  $k_x = \pm 0.7 \mu\text{m}^{-1}$  and  $E = 1.352$  to 1.359 eV, we obtain the photon occupancy  $I_p(k \approx 0)$  after accounting for the independently measured collection efficiency of the optical path (see Methods). As  $I_p(k \approx 0)$  approaches one,  $I_p(k \approx 0)$  shows clearly a superlinear increase



**Fig. 2 | Properties of the heterobilayer and grating cavity.** **a**, An optical microscope image of the WSe<sub>2</sub>/MoSe<sub>2</sub> heterobilayer integrated on a grating cavity. The red square outlines the grating region, and the red circle indicates the laser spot size. Inset, direction of the grating bars. **b**, PL spectrum from the heterobilayer. The sample was pumped with a 633-nm laser at power of 20 μW. The shaded boxes highlight the spectral range of interlayer ( $I_x$ ), MoSe<sub>2</sub> and WSe<sub>2</sub> exciton emission. **c**, TE-polarized along-bar, angle-resolved, simulated (left and overlaid crosses in the right) and measured (right) reflectance spectrum. Inset, line-cut of the normalized reflectance spectrum around  $k_x \approx 1.7 \mu\text{m}^{-1}$  (blue trace); the red line is a fit to the cavity mode. The star symbol marks the peak of the fitted cavity mode. **d**, PL spectrum (blue dots) near  $k_x \approx 0$ . The pump was on resonance with WSe<sub>2</sub> at a pump power of 0.1 μW. The red line is a Lorentzian fit, with a fitted linewidth of 2.4 meV.



**Fig. 3 | Spectral properties of the interlayer exciton laser.** **a**, Angle-resolved micro-PL spectra for the along-bar TE direction at  $P = 0.6 \mu\text{W}$  with overlaid simulated empty cavity (crosses) and cavity with bilayer (stars) dispersions. **b**, The photon occupancy (red) and linewidth (blue) of the TE emission versus input pump power. The emission intensity is integrated over  $|k_x| < 0.7 \mu\text{m}^{-1}$ ,  $|k_y| < 0.13 \mu\text{m}^{-1}$  and  $E = 1.352\text{--}1.359 \text{ eV}$ . The dot-dashed line indicates linear dependence, the vertical red line marks  $P_{\text{th}}$ , and the horizontal purple line indicates  $I_p = 1$ . **c**, Angle-resolved micro-PL spectra for the along-bar TM direction at  $P = 10 \mu\text{W}$ . **d**, The pump power dependence of the TM emission photon occupancy (red) and linewidth (blue), integrated over  $|k_x| < 2 \mu\text{m}^{-1}$ ,  $E = 1.340\text{--}1.400 \text{ eV}$  (open symbols) and  $|k_x| < 0.7 \mu\text{m}^{-1}$ ,  $E = 1.352\text{--}1.359 \text{ eV}$  (filled symbols). Integration over  $|k_y|$  is  $0.13 \mu\text{m}^{-1}$ . The error bars on the photon occupancy data include the shot noise and detector read noise. The error bars on the linewidth data correspond to the 95% confidence interval of the Lorentzian fit.

with pump power, consistent with the onset of stimulated emission into the cavity mode (Fig. 3b). The power-dependent PL measurement is reproducible, as shown by a measurement performed on a different day (Extended Data Fig. 5).

The pump power at the threshold of  $I_p(k \approx 0) = 1$  is  $P_{\text{th}} = 0.18 \mu\text{W}$ . Considering the typical absorption efficiency (20%) of monolayer  $\text{WSe}_2$ , we obtain the threshold carrier density  $n_{\text{th}} = 5.7 \times 10^{10} \text{ cm}^{-2}$ , in good agreement with the density required for the transparency condition  $n_{\text{tr}} \approx 8 \times 10^{10} \text{ cm}^{-2}$  (see Methods for calculations of  $n_{\text{tr}}$ ). Far above threshold, the output intensity becomes linear with pump power and does not saturate up to  $P = 28P_{\text{th}}$ , the highest power used for TE measurements. The nonlinear increase of the intensity is reproduced by a simplified rate equation model, as described in Methods (Extended Data Fig. 6).

Accompanying the superlinear increase in the emission intensity at threshold, the linewidth of the emission drops sharply, as shown in Fig. 3b, signifying the increase of temporal coherence. The linewidth at excitation powers below  $\sim 0.05 \mu\text{W}$  may be broader, but our detectors are not sufficiently sensitive to detect the emission. The saturation and slight increase of linewidth with increasing power above threshold may be due to interactions among the carriers and spatial mode competition<sup>23</sup>. The sharp lasing emission decreases in intensity as we increase temperature, but persists up to 70 K, suggesting that lasing may survive at 70 K or higher (Extended Data Fig. 7).

In stark contrast with the TE emission, TM-polarized emission is not coupled to the cavity mode and does not show threshold behaviour. The emission becomes detectable only at high pump powers. An example is shown in Fig. 3c for  $P = 10 \mu\text{W}$ . The emission spreads uniformly in  $k$  over the numerical aperture of our collection optics and over a broad energy range of about 70 meV. With increasing pump power, the total integrated emission intensity increases sublinearly with pump power (Fig. 3d) and is a few times weaker than that of the TE intensity.

When integrated over the same small ranges of  $k$  and  $E$  near the lasing mode, the TE and TM output intensities differ by several orders of magnitude (filled diamonds in Fig. 3b, d). In other words, while the TM emission is suppressed and remains broadly distributed in  $k$  and  $E$ , the TE emission is concentrated in ranges of energy and  $k$  that are one to two orders of magnitude smaller, as a result of stimulated emission.

To confirm the extended coherence expected of a laser with a 2D gain medium, we study the first-order spatial coherence function  $g^{(1)}(\mathbf{r}_1, \mathbf{r}_2)$  defined as follows:

$$g^{(1)}(\mathbf{r}_1, \mathbf{r}_2) = \frac{G^{(1)}(\mathbf{r}_1, \mathbf{r}_2)}{\sqrt{G^{(1)}(\mathbf{r}_1, \mathbf{r}_1)G^{(1)}(\mathbf{r}_2, \mathbf{r}_2)}} \quad (1)$$

Here  $G^{(1)}$  is the first-order correlation function, and corresponds to:

$$G^{(1)}(\mathbf{r}_1, \mathbf{r}_2) = \text{Tr}\{\rho E^{(-)}(\mathbf{r}_1)E^{(+)}(\mathbf{r}_2)\} \quad (2)$$

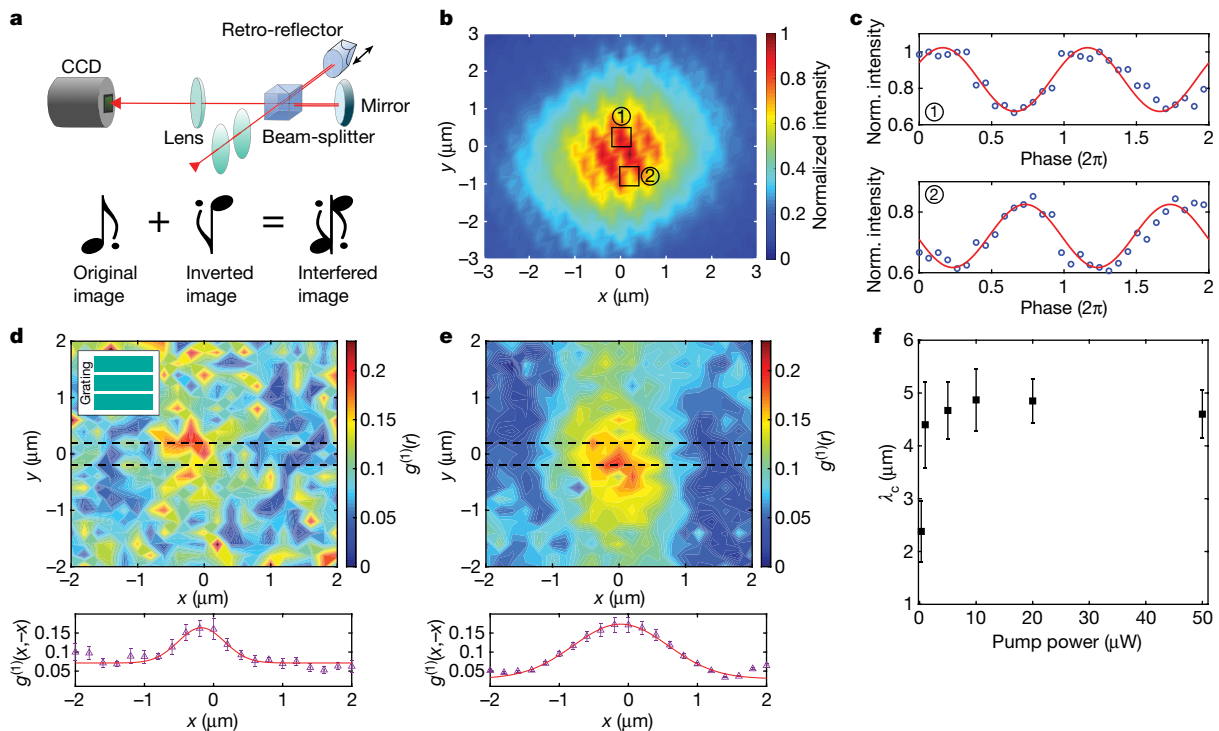
where  $\text{Tr}$  indicates trace,  $\rho$  is the density matrix operator, and  $E^{(+)}$  and  $E^{(-)}$  are field creation and annihilation operators, respectively.

Although spatial coherence properties have been extensively studied in semiconductor photon lasers, exciton–polariton lasers<sup>24</sup> and plasmon lasers<sup>25</sup>, coherence of TMDC lasers has not been studied thus far, making it difficult to rule out localized excitons as a source of lasing. However, the large spatial area of the grating resonator, and the large photon flux above threshold, allow us to investigate the spatial coherence of the interlayer exciton emission. First-order spatial coherence measurements were performed using a continuous wave excitation laser and a retro-reflector Michelson interferometer setup<sup>26</sup>, where an image of the sample interferes with a centro-symmetrically inverted version of itself at the output with an intensity distribution  $I^{\text{int}}(\mathbf{r})$  (Fig. 4a). Because of a small angle difference between the two beams, interference fringes are formed (Fig. 4b) that correspond to slightly varying path length differences  $\frac{2\pi}{\lambda_0}z_0(\mathbf{r})$  at different positions  $\mathbf{r}$  across the images;  $z_0$  is the initial position of the retro-reflector.

Varying the path length of the interferometer,  $I^{\text{int}}(\mathbf{r})$  of each  $\mathbf{r}$  oscillates, with the contrast of the oscillation proportional to the first-order spatial coherence,  $g^{(1)}(\mathbf{r}, -\mathbf{r})$  (Fig. 4c). We thus obtain spatial maps of  $g^{(1)}(\mathbf{r}, -\mathbf{r})$  (see Methods for details). Below threshold, the emission is too weak for  $g^{(1)}(\mathbf{r}, -\mathbf{r})$  measurements. Near threshold, the map is rather noisy without a clear pattern of  $g^{(1)}(\mathbf{r}, -\mathbf{r})$  versus  $\mathbf{r}$  (Fig. 4d, top panel). Above threshold, a clear pattern emerges, showing a high  $g^{(1)}(\mathbf{r}, -\mathbf{r})$  near  $\mathbf{r} \approx 0$  that decays with increasing  $\mathbf{r}$  and extends above the background fluctuations to about twice the laser spot size (Fig. 4e, top panel).

To study the functional dependence of  $g^{(1)}(\mathbf{r}, -\mathbf{r})$ , we average over  $y = \pm 0.2 \mu\text{m}$  and obtain  $g^{(1)}(x, -x)$ . As shown in the bottom panels of Fig. 4d, e, the decay of  $g^{(1)}(x, -x)$  is clearly slower above threshold than below threshold. The plot of  $g^{(1)}(x, -x)$  versus  $x$  is fitted well by a Gaussian function with the standard deviation  $\sigma$  as a fitting parameter. From the fits, we obtain the coherence length  $\lambda_c = \sqrt{2\pi}\sigma$ . As seen in Fig. 4f,  $\lambda_c$  increases abruptly across the threshold, from  $2.38 \mu\text{m}$  near threshold to about  $5 \mu\text{m}$  above threshold, confirming the formation of extended spatial coherence in the laser. Above threshold, the  $\lambda_c$  value remains largely unchanged, possibly limited by the laser spot size and carrier diffusion length. The  $\lambda_c$  value decreases slightly at the highest powers, possibly because of competition of multiple spatial modes in the absence of lateral confinement potentials. We note that the measured  $g^{(1)}(x, -x)$  is much lower than the actual value, owing to difficulty in achieving good alignment.

In conclusion, we have demonstrated a 2D  $\text{WSe}_2$ – $\text{MoSe}_2$  heterobilayer laser on a grating cavity. The injected carrier density at threshold is within an order of magnitude of the estimated transparency condition, suggesting band inversion between the  $\text{MoSe}_2$  conduction band and the  $\text{WSe}_2$  valence band as the gain mechanism. The type-II band alignment, resulting in charge separation and longer exciton lifetimes, may have facilitated the establishment of a population inversion. In addition,



**Fig. 4 | First-order coherence of the interlayer exciton laser.** **a**, Top, schematic of the Michelson interferometer setup. Bottom, illustration of centro-symmetrically interfered images. **b**, Typical interference pattern above  $P_{th}$  (20  $\mu$ W). **c**, Intensity plots of single pixels (labelled as squares 1 and 2 in **b**) as the retro-reflector position is scanned over a phase of  $4\pi$ . **d, e**, Top, maps of  $g^{(1)}(\mathbf{r}, -\mathbf{r})$  near  $P_{th}$  (0.3  $\mu$ W) (**d**) and above  $P_{th}$  (10  $\mu$ W) (**e**). Bottom, horizontal line-cuts of

$g^{(1)}(\mathbf{r}, -\mathbf{r})$  integrated between the dashed lines in the maps above. The red line is the Gaussian fit used to extract the coherence length  $\lambda_c$ . The error bars correspond to the 95% confidence intervals of the sinusoidal fit, such as the ones shown in **c**. Inset in **d**, illustration of the grating bar direction. **f**, The coherence length  $\lambda_c$  versus the pump power. The error bars correspond to the 95% confidence interval of the Gaussian fit, such as the ones shown in **d, e**.

for heterobilayers, moiré lattices are expected<sup>27–31</sup>. By analogy to quantum dot lasers<sup>32</sup>, localization of the interlayer excitons in a moiré lattice may lead to increased phase space density in the lasing mode for the same carrier density, as well as reduced non-radiative loss of the trapped interlayer excitons, enhancing the performance of heterobilayer lasers.

Future studies may clarify the role of moiré lattices in heterobilayer lasers. The present heterobilayer laser could be improved by reducing the inhomogeneous broadening of the gain medium via encapsulation with hexagonal boron nitride, improving the cavity  $Q$ , and reducing mode competition with lateral confinement of the cavity modes. Also, different combinations of van der Waals materials in the heterobilayer could be used to create interlayer exciton lasers of different wavelengths. Using cavities with lateral rotational invariance would allow a valley-polarized interlayer exciton laser to be realized. Finally, electrical tuning of the oscillator strength might allow fast modulation of the laser, and electrical injection could be implemented via atomically thin, bilayer p–n junctions<sup>20,21</sup>. Adiabatic electrical tuning<sup>33</sup> might be used to explore coherent indirect exciton gases.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1779-x>.

- Mak, K. F., Lee, C., Hone, J., Shan, J. & Heinz, T. F. Atomically thin  $\text{MoS}_2$ : a new direct-gap semiconductor. *Phys. Rev. Lett.* **105**, 136805 (2010).
- Splendiani, A. et al. Emerging photoluminescence in monolayer  $\text{MoS}_2$ . *Nano Lett.* **10**, 1271–1275 (2010).
- Geim, A. K. & Grigorieva, I. V. Van der Waals heterostructures. *Nature* **499**, 419–425 (2013).

- Zhang, C. et al. Interlayer couplings, moiré patterns, and 2D electronic superlattices in  $\text{MoS}_2/\text{WSe}_2$  hetero-bilayers. *Sci. Adv.* **3**, e1601459 (2017).
- Fang, H. et al. Strong interlayer coupling in van der Waals heterostructures built from single-layer chalcogenides. *Proc. Natl Acad. Sci. USA* **111**, 6198–6202 (2014).
- Fogler, M. M., Butov, L. V. & Novoselov, K. S. High-temperature superfluidity with indirect excitons in van der Waals heterostructures. *Nat. Commun.* **5**, 4555 (2014).
- Eaton, S. W., Fu, A., Wong, A. B., Ning, C. Z. & Yang, P. Semiconductor nanowire lasers. *Nat. Rev. Mater.* **1**, 16028 (2016).
- Noginov, M. A. et al. Demonstration of a spaser-based nanolaser. *Nature* **460**, 1110–1112 (2009).
- Noda, S. Seeking the ultimate nanolaser. *Science* **314**, 260–261 (2006).
- Wu, S. et al. Monolayer semiconductor nanocavity lasers with ultralow thresholds. *Nature* **520**, 69–72 (2015).
- Ye, Y. et al. Monolayer excitonic laser. *Nat. Photon.* **9**, 733–737 (2015).
- Salehzadeh, O., Djavid, M., Tran, N. H., Shih, I. & Mi, Z. Optically pumped two-dimensional  $\text{MoS}_2$  lasers operating at room-temperature. *Nano Lett.* **15**, 5302–5306 (2015).
- Li, Y. et al. Room-temperature continuous-wave lasing from monolayer molybdenum ditelluride integrated with a silicon nanobeam cavity. *Nat. Nanotechnol.* **12**, 987–992 (2017).
- Shang, J. et al. Room-temperature 2D semiconductor activated vertical-cavity surface-emitting lasers. *Nat. Commun.* **8**, 543 (2017).
- Zhao, L. et al. High-temperature continuous-wave pumped lasing from large-area monolayer semiconductors grown by chemical vapor deposition. *ACS Nano* **12**, 9390–9396 (2018).
- Reeves, L., Wang, Y. & Krauss, T. F. 2D material microcavity light emitters: to lase or not to lase? *Adv. Opt. Mater.* **6**, 1800272 (2018).
- Butov, L. V., Zrenner, A., Abstreiter, G., Böhm, G. & Weimann, G. Condensation of indirect excitons in coupled AlAs/GaAs quantum wells. *Phys. Rev. Lett.* **73**, 304–307 (1994).
- Rigosi, A. F., Hill, H. M., Li, Y., Chernikov, A. & Heinz, T. F. Probing interlayer interactions in transition metal dichalcogenide heterostructures by optical spectroscopy:  $\text{MoS}_2/\text{WS}_2$  and  $\text{MoSe}_2/\text{WSe}_2$ . *Nano Lett.* **15**, 5033–5038 (2015).
- Zhang, L. et al. Highly valley-polarized singlet and triplet interlayer excitons in van der Waals heterostructure. *Phys. Rev. B* **100**, 041402 (2019).
- Lee, C. H. et al. Atomically thin p–n junctions with van der Waals heterointerfaces. *Nat. Nanotechnol.* **9**, 676–681 (2014).
- Ross, J. S. et al. Interlayer exciton optoelectronics in a 2D heterostructure p–n junction. *Nano Lett.* **17**, 638–643 (2017).
- Zhang, L., Gogna, R., Burg, W., Tutuc, E. & Deng, H. Photonic-crystal exciton-polaritons in monolayer semiconductors. *Nat. Commun.* **9**, 713 (2018).
- Spivak, B. & Luryi, S. In *Future Trends in Microelectronics* (eds Spivak, B., Luryi, S. & Zaslavsky, A.) 68–76 (Wiley-Blackwell, 2007).

24. Deng, H., Solomon, G. S., Hey, R., Ploog, K. H. & Yamamoto, Y. Spatial coherence of a polariton condensate. *Phys. Rev. Lett.* **99**, 126403 (2007).
25. Hoang, T. B., Akselrod, G. M., Yang, A., Odom, T. W. & Mikkelsen, M. H. Millimeter-scale spatial coherence from a plasmon laser. *Nano Lett.* **17**, 6690–6695 (2017).
26. Daskalakis, K. S., Maier, S. A. & Kéna-Cohen, S. Spatial coherence and stability in a disordered organic polariton condensate. *Phys. Rev. Lett.* **115**, 035301 (2015).
27. Yu, H., Liu, G.-B., Tang, J., Xu, X. & Yao, W. Moiré excitons: from programmable quantum emitter arrays to spin-orbit-coupled artificial lattices. *Sci. Adv.* **3**, e1701696 (2017).
28. Wu, F., Lovorn, T. & MacDonald, A. H. Theory of optical absorption by interlayer excitons in transition metal dichalcogenide heterobilayers. *Phys. Rev. B* **97**, 035306 (2018).
29. Tran, K. et al. Evidence for moiré excitons in van der Waals heterostructures. *Nature* **567**, 71–75 (2019).
30. Seyler, K. L. et al. Signatures of moiré-trapped valley excitons in MoSe<sub>2</sub>/WSe<sub>2</sub> heterobilayers. *Nature* **567**, 66–70 (2019).
31. Jin, C. et al. Observation of moiré excitons in WSe<sub>2</sub>/WS<sub>2</sub> heterostructure superlattices. *Nature* **567**, 76–80 (2019); correction **569**, E7 (2019).
32. Kirstaedter, N. et al. Low threshold, large T<sub>0</sub> injection laser emission from (InGa)As quantum dots. *Electron. Lett.* **30**, 1416–1417 (1994).
33. Shahnazaryan, V., Kyriienko, O. & Shelykh, I. A. Adiabatic preparation of a cold exciton condensate. *Phys. Rev. B* **91**, 085302 (2015).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019



## Methods

### Sample fabrication

To fabricate the grating cavity, we first grew a SiN film with a SiO<sub>2</sub> buffer on an Si substrate using low-pressure chemical vapour deposition, then patterned it using electron beam lithography and created the grating bars by plasma dry etching. The grating parameters indicated in Fig. 1a are as follows:  $d = 1,475$  nm,  $t = 113$  nm,  $h = 100$  nm,  $\Lambda = 615$  nm and  $g = 50$  nm. The individual WSe<sub>2</sub> and MoSe<sub>2</sub> monolayers were mechanically exfoliated onto a SiO<sub>2</sub> substrate using PDMS (polydimethylsiloxane) polymer. The exfoliated monolayers were stacked into a heterostructure using a high-accuracy rotational alignment method. First, the MoSe<sub>2</sub> was picked up with a PDMS/PPC (polypropylene carbonate) stamp under an optical microscope. Second, the crystal axes of MoSe<sub>2</sub> and WSe<sub>2</sub> were rotationally aligned to be 0° or 60° before stacking. Third, the stacked heterostructure was dropped down onto the grating cavity. Last, the polymer residue was dissolved, and the sample was annealed at 350 °C for a total of 7 h.

### Heterobilayer twist angle

We can verify the twist angle of the heterobilayer aligned under the optical microscope by angle-dependent second-harmonic generation (SHG) measurements. Extended Data Fig. 1 shows an optical microscope image of two different samples and the corresponding angle-dependent SHG measurements. By fitting the SHG pattern with a  $\cos^2(3\theta)$  function, where  $\theta$  is the angle between the armchair direction of the monolayer and the polarization direction of the beam, we can obtain the twist angle. We did not measure the SHG for the heterobilayer before putting it on the grating, but from experience, we have found that the straight edges of exfoliated monolayers reliably correspond to the armchair axis of the crystal. Therefore, we aligned the two straight flake edges under the optical microscope.

### Time-resolved PL

To measure the decay time of the TM emission we used a Hamamatsu streak camera system. The emission was polarization-selected for the TM direction and sent to the streak camera. As shown in Extended Data Fig. 2, a line-cut of the streak camera spectrum was fitted with a bi-exponential function to determine the lifetime. The fitted lifetime is around 2 ns.

### PL mapping of the heterobilayer device

The spatially resolved PL mapping of the heterobilayer device is shown in Extended Data Fig. 3. For this measurement, we use a 633-nm continuous wave (CW) laser and selected TE polarization for excitation. The sample is mounted on an Attocube ANC 300 piezo stage and scanned as needed. Spectral band-pass filters are used to select the emission from bilayer, WSe<sub>2</sub> and MoSe<sub>2</sub> regions.

### Measurements of lasing characteristics

Extended Data Fig. 8 shows the schematic of the optical setup used for the angle-resolved PL reflection and the coherence measurements of the heterobilayer laser device. The sample was cooled to 5 K using a Montana Instruments Fusion 2 cryostat. Fourier-space imaging was used to measure angle-resolved reflection and micro-PL of the device. For reflection, a tungsten halogen lamp was used. For micro-PL, a pulsed Ti:sapphire laser (80 MHz repetition rate, 150 fs pulse width) near-resonant with the WSe<sub>2</sub> A-exciton (1.7 eV) was used to excite the sample. The emission was collected using a 0.42 NA objective lens, passed through a long-pass filter to filter out the excitation laser and a linear polarizer to selectively measure TE and TM modes, and sent to a Princeton Instruments spectrometer with a measured spectral resolution of 0.3 nm. The entrance slit of the spectrometer is aligned along the y direction. The slit width of 100  $\mu$ m corresponds to a range  $|k_y| < 0.13 \mu\text{m}^{-1}$ . The NA of the collection optics corresponds to a range of  $|k_x| < 2 \mu\text{m}^{-1}$ .

Spatial coherence measurement was performed using a retro-reflector Michelson Interferometer setup as shown in Extended Data Fig. 8. Emission rid of scattered pump laser light was sent to a 50:50 beam splitter which divided the light into two paths, the mirror path and the retro-reflector path. In order to change the time difference ( $\tau$ ) between the two paths, the retro-reflector is mounted on a stepper motor which can have a step size as small as 50 nm (about 0.167 fs). The interference pattern was collected by a Princeton Instruments eXcelon charge-coupled camera, with intensities described by:

$$I^{\text{int}}(\mathbf{r}) = I(\mathbf{r}) + I(-\mathbf{r}) + 2\sqrt{I(\mathbf{r})I(-\mathbf{r})}g^{(1)}(\mathbf{r}, -\mathbf{r})\sin\left(\frac{2\pi}{\lambda_0}(z - z_0)\right) \quad (3)$$

Here  $I(\mathbf{r})$  and  $I(-\mathbf{r})$  are intensities from the mirror and retro-reflector arms, respectively, and are measured by blocking one of the arms of the interferometer,  $z$  is the position of the retro-reflector,  $g^{(1)}(\mathbf{r}, -\mathbf{r})$  is the first-order spatial coherence for two positions separated by  $2r$ , and is proportional to the visibility of the interference fringe. To obtain the visibility or  $g^{(1)}(\mathbf{r}, -\mathbf{r})$ , we scan the position  $z$  of the retro-reflector and record the sinusoidal oscillation of  $I^{\text{int}}(\mathbf{r})$  versus  $z$  at each  $\mathbf{r}$ , as shown in Fig. 4c.

The emission from our ultra-compact device is necessarily weak and the detector efficiency at 1.35 eV is poor, hence it is difficult to simultaneously achieve good time and spatial overlap of the interference signal with the asymmetric interferometer. With a symmetric interferometer built with two retro-reflectors, the alignment is much less sensitive to slight changes in the incident beam; therefore we are able to achieve good alignment of the signal path by using an auxiliary alignment laser and obtain visibility for  $g^{(1)}(\tau = 0)$  close to 0.8 (Extended Data Fig. 9).

### Photon number

Photon occupancy per pulse ( $I_p(k \approx 0)$ ) was estimated from the total count rate on the detector,  $n_c$ . The total integration time for angle-resolved spectra was 90 s. The two values are related by:  $I_p = \eta \rho f n_c$  ( $k \approx 0$ ). Here  $\eta \approx 10^{-7}$  is the total detection efficiency of the setup, which is independently calibrated by replacing the sample with a laser-coupled single-mode fibre,  $\rho \approx 1$  is the number of  $k$ -space modes within the integrated region, and  $f = 80$  MHz is the repetition rate of the pump laser.

### Transparency condition

The transparency condition is defined as the number of carriers required for the energy difference between the quasi-Fermi levels in the conduction ( $E_{F,c}$ ) and valence ( $E_{F,v}$ ) bands to equal the lasing energy ( $E_{F,c} + E_{F,v} = 0$ ). The quasi-Fermi levels are determined by the electron density:

$$N_c = N_c \int_0^\infty \frac{1}{1 + \exp[\varepsilon_c - \varepsilon_{F,c}]} d\varepsilon_c \quad (4)$$

Here,  $\varepsilon_c = E_c/k_B T$  and  $\varepsilon_{F,c} = E_{F,c}/k_B T$ ,  $E_c$  is the conduction band edge,  $k_B$  is the Boltzmann constant and  $T$  is temperature.  $N_c$  is the effective density of states in two dimensions for electrons with an effective mass  $m_e^*$ :

$$N_c = \frac{m_e^* k_B T}{\hbar^2 \pi} \quad (5)$$

Solving equation. (4) for  $\varepsilon_{F,c}$  we obtain:

$$\varepsilon_{F,c} = \ln \left[ \exp \left( \frac{N_e}{N_c} \right) - 1 \right] \quad (6)$$

The equation for valence band Fermi energy and hole carrier density can be written in a similar way. Assuming  $n = N_e = N_h$  and using the effective masses of K-valley electron and holes given in ref. <sup>34</sup>, we solve for the carrier density that satisfies the transparency condition and obtain

$n_{tr} = 8 \times 10^{10} \text{ cm}^{-2}$ , which is in good agreement with the threshold carrier density  $n_{th} = 5.7 \times 10^{10} \text{ cm}^{-2}$ .

## Simplified rate equation model of the laser

We use the following rate equations to describe the time evolution of interlayer exciton density  $N$  and the photon density  $S$  in the lasing mode:

$$\frac{dN}{dt} = \frac{\eta P}{\hbar \omega V_a} - \frac{(1 - \beta_0)N}{\tau_{sp}} - \frac{F\beta_0 N}{\tau_{sp}} - av_g(N - N_{tr})S \quad (7)$$

$$\frac{dS}{dt} = \Gamma \frac{F\beta_0 N}{\tau_{sp}} + \Gamma av_g(N - N_{tr})S - \frac{S}{\tau_p} \quad (8)$$

Parameters in the equation are listed in Extended Data Table 1. In the heterobilayer system, electron and hole transfer takes place on the subpicosecond timescale, much shorter than the interlayer exciton lifetime. Therefore, we consider a pulsed pump  $P$  that creates an initial carrier population of  $N(t=0) \propto P$ . The photon number  $I_p$  is proportional to the photon density, and  $I_p = 1$  at threshold. The simulated curve of  $I_p$  versus pump power matched the experiment well, as shown in Extended Data Fig. 6.

## Data availability

Data that support the findings of this study are available from the corresponding author on reasonable request.

34. Jin, Z., Li, X., Mullen, J. T. & Kim, K. W. Intrinsic transport properties of electrons and holes in monolayer transition-metal dichalcogenides. *Phys. Rev. B* **90**, 045422 (2014).

**Acknowledgements** We acknowledge support from the Army Research Office under awards W911NF-17-1-0312. H.D. acknowledges support from the Asian Office of Aerospace R&D under awards FA2386-18-1-4086. E.T. acknowledges support from Intel Corp., and the Welch Foundation grant F-2018-20190330.

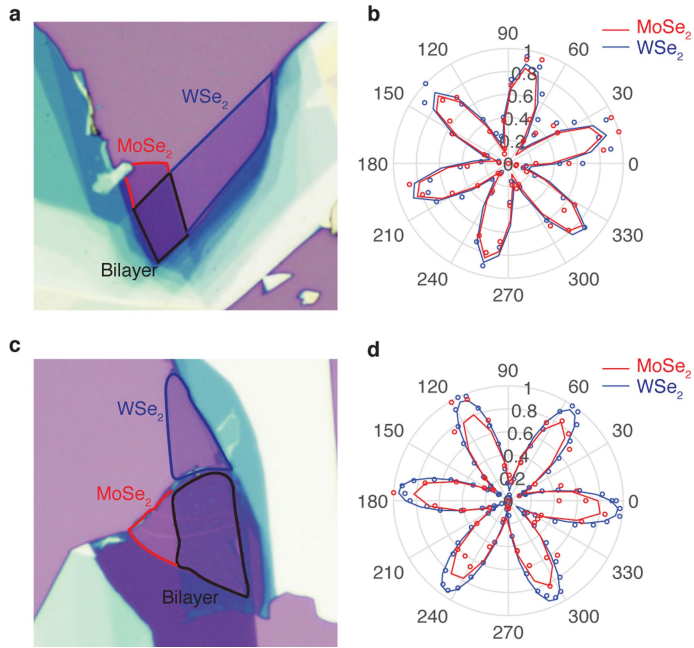
**Author contributions** E.Y.P. performed the measurements of the indirect exciton laser. L.Z. designed, fabricated and characterized the grating cavity. L.Z. exfoliated and characterized the WSe<sub>2</sub> and MoSe<sub>2</sub> monolayers with help from E.Y.P. G.W.B. made the rotationally aligned structure. R.G., E.Y.P. and L.Z. performed simulations of the device. H.D. and E.T. supervised the project. E.Y.P., L.Z. and H.D. performed data analysis and wrote the manuscript with input from all authors.

**Competing interests** The authors declare no competing interests.

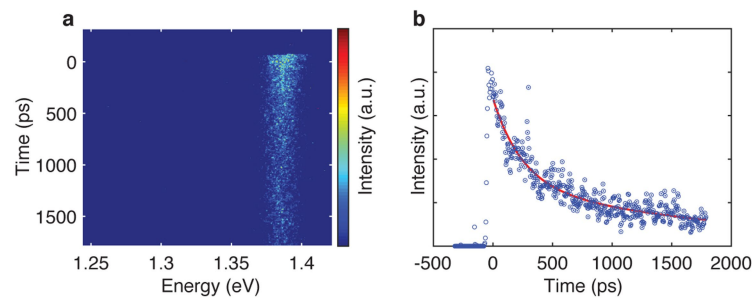
## Additional information

**Correspondence and requests for materials** should be addressed to H.D.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

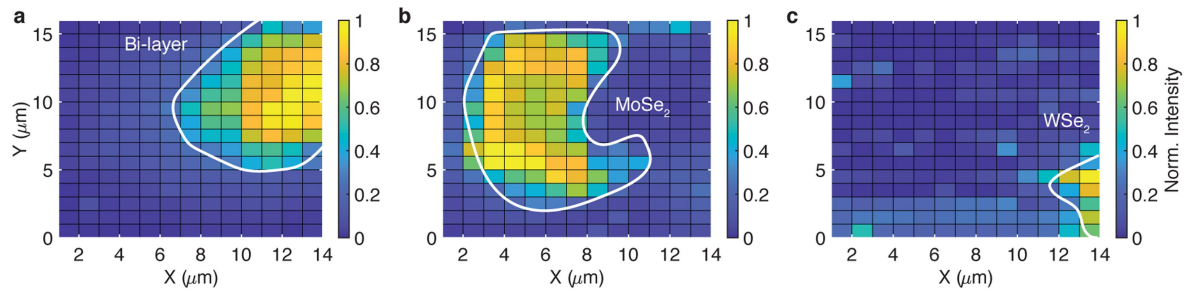


**Extended Data Fig. 1 | Heterobilayer twist angle. a, b,** Optical image (a) and angle-resolved SHG measurements (b; open circles) of WSe<sub>2</sub>/MoSe<sub>2</sub> heterobilayers. **c, d,** As **a, b** but for a different sample. The field of view of the optical images is around 60 μm. Solid lines in **b, d**, are fits by a  $\cos^2(3\theta)$  function, which give relative twist angles of  $0.22^\circ \pm 1.78^\circ$  for **b**, and  $0.34^\circ \pm 1.5^\circ$  for **d**.

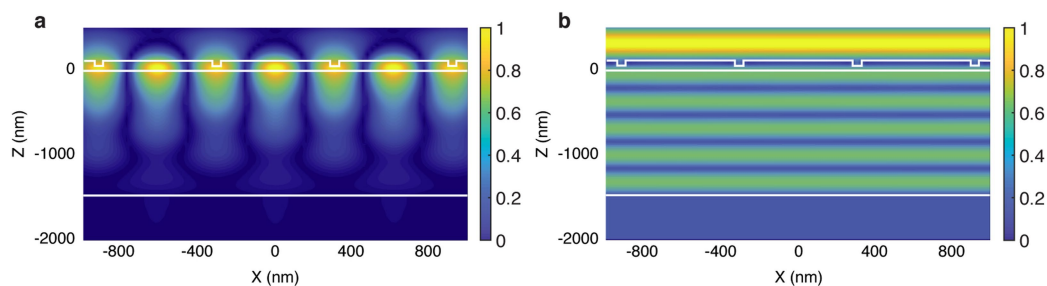


**Extended Data Fig. 2 | Interlayer exciton lifetime.** **a**, Time-resolved PL spectrum for TM emission. **b**, Line-cut of **a** near 1.38 eV. Red line is a bi-exponential fit to the data, with a fitted lifetime of 2 ns.



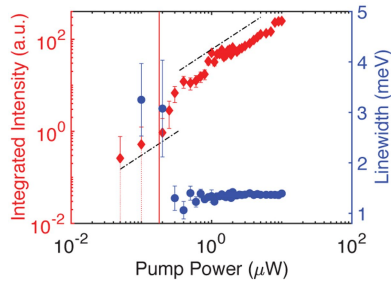


**Extended Data Fig. 3 | Spatial mapping of PL.** The normalized intensity of PL from the device is shown as a function of position. Spectral filters centred around their respective exciton peak energies were applied for each image. The white contours mark the regions of heterobilayer (a), MoSe<sub>2</sub> (b) and WSe<sub>2</sub> (c).

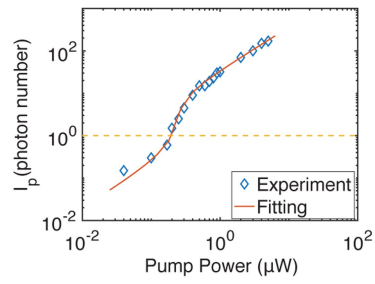


**Extended Data Fig. 4 | Electric field profiles.** Shown are simulated normalized electric field profiles as a function of position near the centre of a grating cavity with lateral dimensions of  $100\ \mu\text{m} \times 100\ \mu\text{m}$ . **a**, TE-polarized light at the cavity resonance at  $k=0$ , showing strong field enhancement in the grating layer including at its surface where the heterobilayer is placed. **b**, TM-polarized light

at the same wavelength as **a**, showing negligible cavity effects. White lines outline different layers of the grating cavity. The corresponding enhancement of the exciton radiative decay rate, or the Purcell factor, is calculated to be around 2.4.

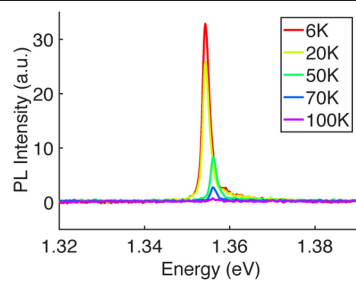


**Extended Data Fig. 5 | Power-dependence reproducibility.** The photon occupancy (red) and linewidth (blue) of TE emission from the heterobilayer versus input pump power, similar to that shown in Fig. 3b of the main text but measured on a different day to show the reproducibility of the device. The error bars on the photon occupancy data include the shot noise and detector read noise. The error bars on the linewidth data correspond to the 95% confidence interval of the Lorentzian fit.

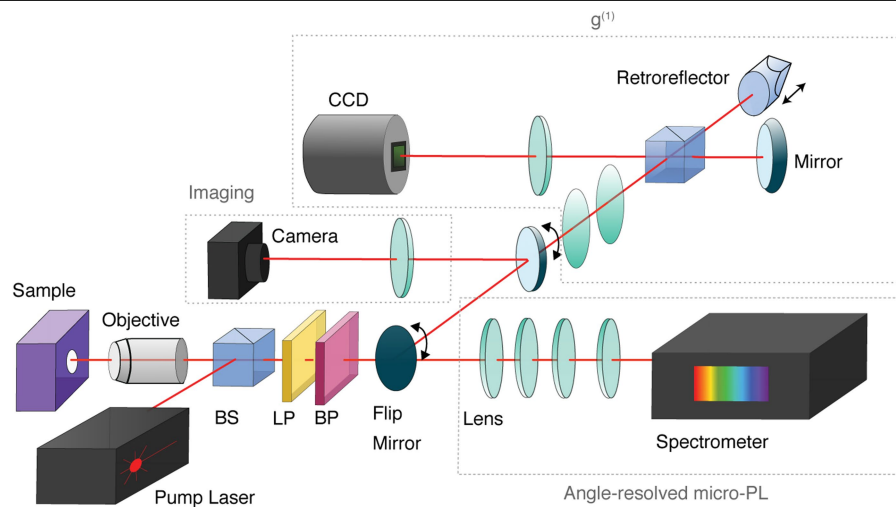


**Extended Data Fig. 6 | Rate equation fitting.** The log-log plot of photon occupancy versus pump power. The diamonds represent measured data shown in Fig. 3b of the main text, and the solid line is a rate-equation fitting. Details of the rate equation simulation are described in Methods.

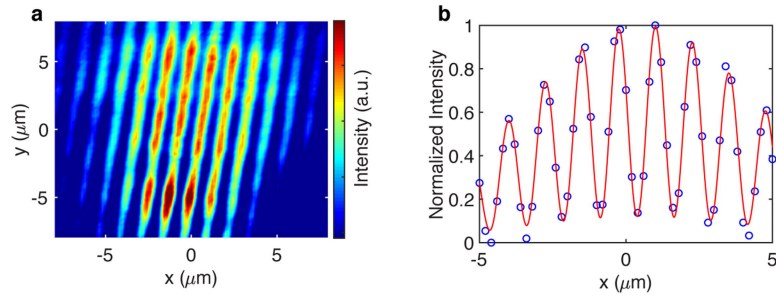




**Extended Data Fig. 7 | Temperature dependence.** Temperature-dependent real space PL spectra of the sample studied in the main text.



**Extended Data Fig. 8 | Experimental setup.** Schematic diagram of the optical experimental setup as described in Methods. BS, beam splitter; LP, long-pass filter; BP, band-pass filter; CCD, charge-coupled device.



**Extended Data Fig. 9 | Temporal coherence.** Shown are temporal coherence interference fringes at  $g^{(1)}(\tau=0)$  measured using a two-retro-reflector Michelson interferometer under CW excitation above threshold.

**a**, Interferogram image at  $\tau=0$ . **b**, Horizontal line-cut of **a** around  $y=1.5\ \mu\text{m}$ . The red line is a fit to the Gaussian pump beam profile modulated by a cosine function. Here  $g^{(1)}(\tau=0)=0.78$ .

Extended Data Table 1 | Rate equation fitting parameters

Parameter	Definition	Value
$F$	Purcell factor	2.35
$\Gamma$	Confinement factor	0.0208
$V_a$	Carrier injection volume	$1.4 \times 10^{-3} \mu\text{m}^3$
$\tau_{sp}$	Spontaneous emission lifetime	2 ns
$\tau_p$	Photon lifetime	0.3 ps
$\eta$	Absorption efficiency	20 %
$N_{tr}$	Transparency density	$8 \times 10^{-10} \text{cm}^{-2}$
$\beta_0$	Spontaneous emission factor	0.046
$a$	Absorption cross section	$1.9 \times 10^{-14} \text{cm}^{-2}$
$v_g$	Group velocity	$1.5 \times 10^8 \text{m/s}$

Definitions and values of the parameters used in the rate equation simulation. The values of  $F$ ,  $\Gamma$  and  $V_a$  are estimates for our structure;  $N_{tr}$  is estimated in Methods; measured values are used for  $\tau_{sp}$ ,  $\tau_p$  and  $\eta$ ;  $\beta_0$  and  $a$  are fitting parameters. See Methods for definition of the rate equation using these parameters.



# Thermoelectric performance of a metastable thin-film Heusler alloy

<https://doi.org/10.1038/s41586-019-1751-9>

Received: 30 January 2018

Accepted: 22 August 2019

Published online: 13 November 2019

B. Hinterleitner<sup>1,2,9</sup>, I. Knapp<sup>1,2,9</sup>, M. Poner<sup>1,2,9</sup>, Yongpeng Shi<sup>3,4,9</sup>, H. Müller<sup>1</sup>, G. Eguchi<sup>1</sup>, C. Eisenmenger-Sittner<sup>1</sup>, M. Stöger-Pollach<sup>1,5</sup>, Y. Kakefuda<sup>6</sup>, N. Kawamoto<sup>6</sup>, Q. Guo<sup>6,7</sup>, T. Baba<sup>6,7</sup>, T. Mori<sup>6,7,8</sup>, Sami Ullah<sup>3</sup>, Xing-Qiu Chen<sup>3,4</sup> & E. Bauer<sup>1,2,9\*</sup>

Thermoelectric materials transform a thermal gradient into electricity. The efficiency of this process relies on three material-dependent parameters: the Seebeck coefficient, the electrical resistivity and the thermal conductivity, summarized in the thermoelectric figure of merit. A large figure of merit is beneficial for potential applications such as thermoelectric generators. Here we report the thermal and electronic properties of thin-film Heusler alloys based on  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  prepared by magnetron sputtering. Density functional theory calculations suggest that the thin films are metastable states, and measurements of the power factor—the ratio of the Seebeck coefficient squared divided by the electrical resistivity—suggest a high intrinsic figure of merit for these thin films. This may arise from a large differential density of states at the Fermi level and a Weyl-like electron dispersion close to the Fermi level, which indicates a high mobility of charge carriers owing to linear crossing in the electronic bands.

Thermoelectric devices are able to directly convert thermal energy into electrical energy. The efficiency of this process depends on the temperature difference between the hot ( $T_h$ ) and the cold ( $T_c$ ) sides of the device (defining the Carnot efficiency) and the performance of the thermoelectric material, as expressed by the thermoelectric figure of merit  $ZT = \frac{S^2}{\rho\lambda}T$ , where  $S$ ,  $\rho$  and  $\lambda$  are the Seebeck coefficient, the electrical resistivity and the thermal conductivity, respectively, and  $T$  is the temperature at which the thermoelectric properties are measured. To improve the thermoelectric performance of a certain material, the power factor,  $PF = S^2/\rho$ , must be increased and the thermal conductivity,  $\lambda = \lambda_e + \lambda_{ph}$ , must be reduced<sup>1</sup> ( $\lambda_e$  and  $\lambda_{ph}$  denote the electronic and phononic contributions to  $\lambda$ , respectively).

The three individual physical properties constituting the figure of merit are not independent from each other. Therefore, improving one without causing another to deteriorate is difficult or impossible.  $\lambda_{ph}(T)$  is the only quantity that can be changed freely without influencing the others. Thus, the most promising way to improve  $ZT$  is the reduction of dimensions and dimensionalities<sup>2</sup>. In this regard, a vast quantity of work has been done to obtain and study thermoelectric materials with appropriate length scales, down to a few nanometres<sup>3</sup>.

The focus of the present study is thin-film full-Heusler alloys deposited on Si wafers. Besides the expectation that the thermoelectric properties will be enhanced, thin films can also be a basis for applications in fields such as microelectronics.

Half- and full-Heusler systems are useful thermoelectric materials owing to their reasonably large  $PF$  and  $ZT$  values and the modest costs of the materials, as well as their chemical and mechanical long-term

stability<sup>4,5</sup>. Whereas half-Heusler alloys can be described as XYZ ternary compounds, full-Heusler systems have a composition of the form  $\text{X}_2\text{YZ}$ , where X and Y are (in general) transition-metal elements and Z is a main-group element. Depending on the specific composition, various ground states are possible, including semiconducting states with different gaps in their electronic density of states (DOS).

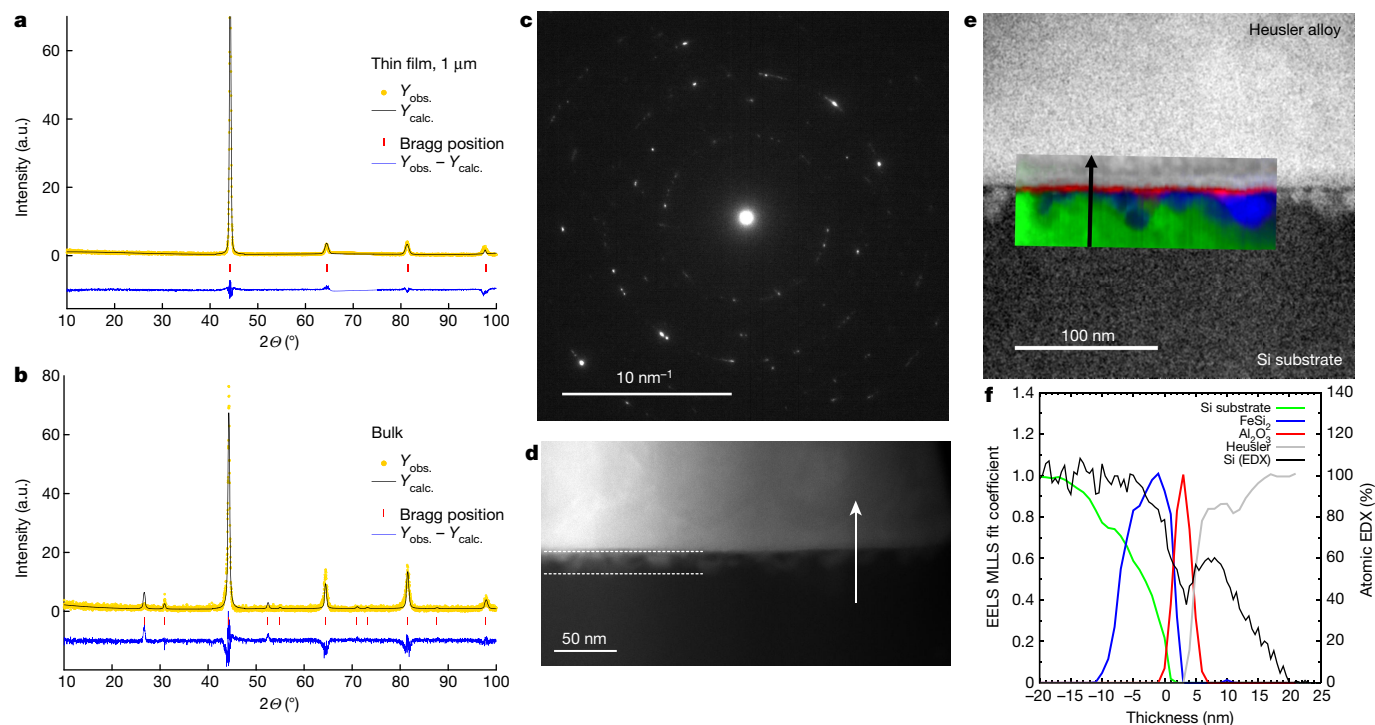
We start by studying  $\text{Fe}_2\text{VAl}$ . Although three metallic elements constitute a ternary system,  $\text{Fe}_2\text{VAl}$  exhibits a (pseudo-)gap in its electronic DOS near the Fermi energy  $E_F$ , located near the valence-band edge, where some residual states lie<sup>6</sup>. Elemental substitutions in  $\text{Fe}_2\text{VAl}$  enable us to tailor the electronic properties by modifying the electronic DOS close to  $E_F$  and creating scattering centres to minimize the lattice thermal conductivity. Following this strategy, we prepared a series of compounds of the form  $\text{Fe}_2\text{V}_{1-x}\text{W}_x\text{Al}$ .

The material studied here reveals a high thermoelectric performance, well above any numbers reported in the literature so far<sup>7–11</sup>. However, the metastable state of these films (as demonstrated below), despite excellent intrinsic thermoelectric properties, might produce some challenges in fabricating useful thermoelectric devices.

## Crystal structure

Figure 1a, b summarizes the X-ray diffraction pattern obtained for  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  from both the thin-film and bulk samples. The corresponding Rietveld refinements, that is, the use of a nonlinear least-squares fit to minimize the differences between the entire set of observed X-ray peak intensities and the peaks calculated from a crystal

<sup>1</sup>Institute of Solid State Physics, Technische Universität Wien, Vienna, Austria. <sup>2</sup>Christian Doppler Laboratory for Thermoelectricity, Technische Universität Wien, Vienna, Austria. <sup>3</sup>Shenyang National Laboratory for Materials Science, Institute of Metal Research, Chinese Academy of Sciences, Shenyang, China. <sup>4</sup>School of Materials Science and Engineering, University of Science and Technology of China, Shenyang, China. <sup>5</sup>University Service Centre for Transmission Electron Microscopy, Technische Universität Wien, Vienna, Austria. <sup>6</sup>International Center for Materials Nanoarchitectonics (WPI-MANA), National Institute for Materials Science (NIMS), Tsukuba, Japan. <sup>7</sup>Center for Functional Sensor & Actuator (CFSN), National Institute for Materials Science (NIMS), Tsukuba, Japan. <sup>8</sup>University of Tsukuba, Tsukuba, Japan. <sup>9</sup>These authors contributed equally: B. Hinterleitner, I. Knapp, M. Poner, Yongpeng Shi, E. Bauer. \*e-mail: [bauer@ifp.tuwien.ac.at](mailto:bauer@ifp.tuwien.ac.at)



**Fig. 1 | Structure and phases of  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  and  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$ -Si substrate from X-ray diffraction and electron microscopy. a, b, X-ray diffraction of  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  ( $Y_{\text{obs.}}$ , yellow circles), alongside results from a Rietveld refinement ( $Y_{\text{calc.}}$ , black), the difference between experimental and model data (blue) and the respective Bragg positions (red) for thin-film (a) and bulk (b) material. c, Electron diffraction pattern of the Heusler thin film. d, Z-contrast image of the Heusler material-substrate interface, showing the formation of  $\text{Fe}_2\text{Si}$  crystals inside the substrate. The dashed lines denote the width of the  $\text{Fe}_2\text{Si}$**

model, are shown (solid lines), together with the respective Bragg positions (vertical lines) and the difference between the fit and the experimental data.

There are distinct differences among the X-ray patterns between the bulk and the thin-film sample. Although the bulk sample essentially follows the predictions of the standard full-Heusler system (face-centred cubic (fcc),  $\text{Cu}_2\text{MnAl}$ -type structure, space group  $Fm\bar{3}m$ ), the film is characterized by the absence of a number of Bragg peaks, which indicates that it belongs to the more simple W-type structure, as a subgroup of full-Heusler systems (body-centred cubic (bcc), space group  $Im\bar{3}m$ ). Although in the fcc structure Fe is located at the (8c) sites, V and W are located at (4b) and Al lies at (4a). If some anti-site occupation on the (4a) and (4c) sites occurs, the X-ray intensity of the (111) peak weakens and even vanishes for a random disorder of V and Al. A CsCl-type structure results. The absence of the (111) and (200) X-ray peaks in the sputtered and heat-treated film is evidence of further intermixing of atoms on the various lattice sites. In the thin-film sample, a W-type structure is realized and the system changes from fcc to bcc<sup>4</sup>. All atoms are equally distributed on the (2a) sites of this structure. Formally, the lattice parameter changes from  $a = 5.7864 \text{ \AA}$  to  $a = 2.8977 \text{ \AA}$ .

### The film-substrate interface

To derive detailed information in terms of composition and morphology about the thin film (thickness,  $1 \mu\text{m}$ ), the substrate (thickness,  $280 \mu\text{m}$ ), and the interface in between, we have carried out various electron-microscopy-based investigations on  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Si}$  sputtered on Si.

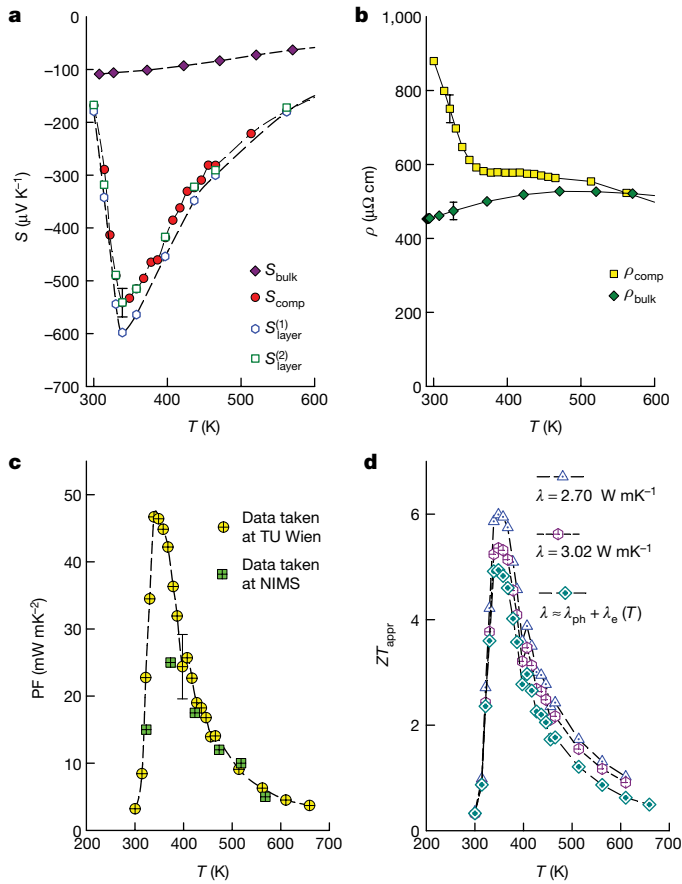
The electron diffraction pattern in Fig. 1c shows the polycrystalline nature of the Heusler thin film. Because the thin film is deposited on a

rough surface, no preferential orientation can be observed. In Fig. 1d a scanning transmission electron microscopy image of the interface is shown, using Z-contrast conditions. A diffusion zone with thickness of about  $20 \text{ nm}$  can be seen where  $\text{Fe}_2\text{Si}$  forms. This interlayer extends into the Si substrate, but does not appear as a compact  $\text{Fe}_2\text{Si}$  layer. Instead,  $\text{Fe}_2\text{Si}$  forms as very weakly connected islands, typically a few tens of nanometres in size.

This layer is the result of diffusion of the deposited film at elevated substrate temperatures and of diffusion during heat treatment of the sample.  $\text{Fe}_2\text{Si}$ , the primary product formed in this interface, is a metallic ferromagnet, crystallizing either in a tetragonal or hexagonal crystal structure<sup>12,13</sup>. The metallic nature of this material, as is clear from band structure calculations, will probably generate only moderate thermopower values. In any case, we expect its contribution to the total electronic and thermal transport to be restricted to a few per cent of the overall measured effects, because this thin  $20\text{-nm}$  interlayer corresponds to only 2% of the active thermoelectric Heusler alloy.

We then used electron-energy-loss spectroscopy (EELS) to analyse the interface in detail. We found that an alumina layer with a thickness of  $5 \text{ nm}$  is also formed at the interface. Figure 1e gives a compositional map created from EELS measurements, showing Si in green,  $\text{Fe}_2\text{Si}$  in blue,  $\text{Al}_2\text{O}_3$  in red and the Heusler phase in light grey.

A combined EELS and energy-dispersive X-ray spectroscopy (EDX) profile, displayed in Fig. 1f, demonstrates that Si diffuses into the Heusler thin film as well. Whereas the EELS data were fitted using a multiple linear least-squares fit routine, in Fig. 1f the EDX data for Si were normalized to the maximum peak height to give a better picture of the Si distribution at the interface. It can be seen that the Si diffused up to



**Fig. 2 | Temperature-dependent transport and thermoelectric properties of thin-film  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$ .** **a, b,** The temperature-dependent Seebeck coefficient (**a**) and electrical resistivity (**b**) of the entire composite (layer, interface and substrate), together with the thin-film value of  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$ , obtained from equation (2).  $S_{\text{layer}}^{(1)}$  and  $S_{\text{layer}}^{(2)}$  are the deduced Seebeck data with and without  $\text{Fe}_2\text{Si}$  as the interface, respectively. The corresponding data for the bulk material are added for comparison. **c, d,** The temperature-dependent power factor (**c**) and approximated figure of merit (**d**). PF values shown here refer to  $S_{\text{comp}}$  of Fig. 2a. The size of the error bar in **c** results from an estimated error of 5% for both the Seebeck (**a**) and resistivity (**b**) data.  $ZT_{\text{appr}}$  is evaluated to a first approximation using the room-temperature (25 °C) thermal conductivity ( $\lambda_{\text{RT}}^{\text{diff}} = 2.70 \text{ W m}^{-1} \text{ K}^{-1}$ ;  $\lambda_{\text{RT}}^{\text{eff}} = 3.02 \text{ W m}^{-1} \text{ K}^{-1}$ ). From the Wiedemann–Franz law, a temperature-dependent thermal conductivity is estimated, keeping  $\lambda_{\text{ph}}$  constant. The respective  $ZT_{\text{appr}}$  data are indicated by open diamonds.

20 nm into the Heusler thin film. The energy loss near edge structures (ELNES) (see Extended Data Fig. 1) was used to fit the phase distribution shown in Fig. 1f.

## Thermoelectric properties

To define the thermoelectric properties of the  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  thin film, electrical resistivity, thermal conductivity and Seebeck measurements were carried out at Technische Universität (TU) Wien, Vienna, Austria. Figure 2a, b shows the temperature-dependent Seebeck coefficient,  $S$ , and the electrical resistivity,  $\rho$ , respectively, of  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  thin film annealed at 450 °C for one week.  $\rho(T)$  of this Heusler alloy is characterized by semiconductor-like resistivity behaviour, with decreasing resistivity as the temperature increases. Absolute  $\rho(T)$  values correspond fairly well with those derived for bulk materials with the same concentration of W. In addition, the overall resistivity of the thin film at lower temperatures is smaller by a factor of 2–3 compared to the resistivity of the starting material  $\text{Fe}_2\text{VAl}$ . In agreement with density

functional theory (DFT) calculations, the V–W substitution shifts the Fermi energy towards the band edge of the conduction band, with an increased charge carrier density compared to  $\text{Fe}_2\text{VAl}$ .

The Seebeck coefficient  $S(T)$  of  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  exhibits very large negative values, indicating that electrons are the majority charge carriers in this system. This conclusion is supported by Hall effect data taken at high temperatures (details are discussed in Methods.) The largest values of the Seebeck coefficient occur at moderately high temperatures, similar to archetypal Bi–Te compounds.

To corroborate these results, additional studies of transport properties were done at the National Institute for Materials Science (NIMS), Tsukuba, Japan, using a device similar to that used at TU Wien (ZEM-3, ULVAC), and a different device (ZEM-2, ULVAC). These measurements confirm the initial results. Physical property measurement system (PPMS, Quantum Design) measurements based on a different data acquisition principle yielded some dissimilar results; the Seebeck coefficient obtained using that measuring technique, revealed—in part—even larger absolute values.

To get a closer look and understanding of individual contributions of the active thermoelectric layer, the interlayer and the Si substrate, a parallel-conductance model is assumed, in which the constituent parts of the composite sample contribute individually to the observed effect. In terms of the electrical resistivity, the well known parallel-resistance model can be applied:

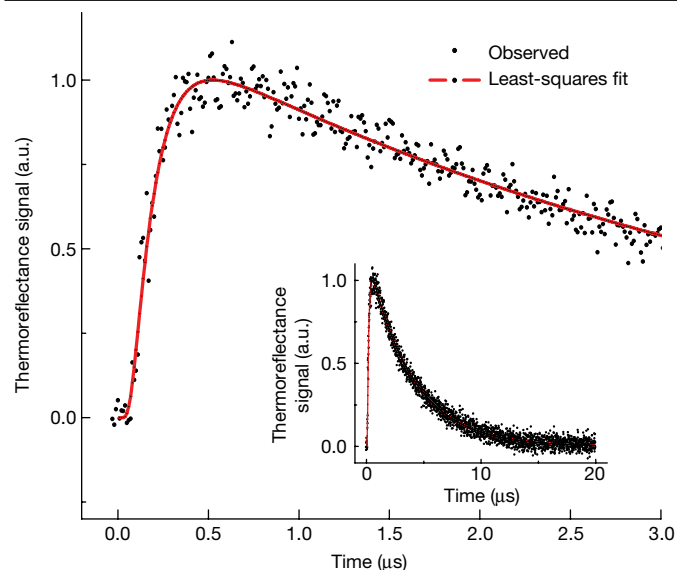
$$\frac{1}{R} = \sum_i \frac{1}{R_i} \quad (1)$$

Similarly, the Seebeck coefficient can be understood from

$$S_{\text{total}} = \frac{\sum_i S_i \sigma_i d_i}{\sum_i \sigma_i d_i} \quad (2)$$

where  $i = \{\text{layer, int, sub}\}$  denotes the active thermoelectric layer, the interface and the substrate, respectively. Equation (2) indicates that in such a distinct structure the individual Seebeck coefficients  $S_i$  contribute to the total measured coefficient in a weighted manner, corresponding to the electrical conductivities  $\sigma_i = 1/\rho_i$  and the slab thicknesses  $d_i$  of the slabs involved. To ensure reliable results, the active thermoelectric layer was removed mechanically from the substrate by polishing, and then re-measured. The resistivities, which are 4–5 orders of magnitude larger than that of the Heusler film, were obtained, together with large negative Seebeck values (see Extended Data Fig. 3). In the absence of available data, a very large Seebeck effect is assumed for the interface (predominantly  $\text{Fe}_2\text{Si}$ ;  $S_{\text{int}} \approx 100 \mu\text{V K}^{-1}$ ), together with a very low electrical resistivity ( $\rho_{\text{int}} \approx 100 \mu\Omega \text{ cm}$ ; this may be considerably underestimated, since the  $\text{Fe}_2\text{Si}$  interface layer is not a continuous film; instead, it seems to be made up of small  $\text{Fe}_2\text{Si}$  islands, weakly connected and arranged along the interface (see Fig. 1d). Based on the above data and assumptions, the Seebeck coefficient of the active layer,  $S_{\text{layer}}$ , can first be derived from equation (2). Obtained in this way,  $S_{\text{layer}}^{(1)}(T)$  is added to Fig. 2a. As a second step, the  $\text{Fe}_2\text{Si}$  interface is neglected in the analysis, because the condition of percolation might not be fulfilled. Data from this procedure are labelled as  $S_{\text{layer}}^{(2)}$  in Fig. 2a. We note that in this second case,  $S_{\text{layer}}^{(2)}(T)$  is almost the same as the total measured value,  $S_{\text{comp}}(T)$ . The intrinsic nature of the very large thermopower values of thin-film  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  can be inferred from a direct comparison of the thin-film data evaluated here and those of the pure substrate. As shown in Extended Data Fig. 3, the pure substrate exhibits even larger values; twice as large in a temperature range from 500 to 800 K.

There are only minor changes compared to the experimentally deduced data owing to the very thin interface (compared to the active layer and the thick substrate), and the very large resistivity values in the lower temperature range of the substrate. These data indicate that a



**Fig. 3 | Time-dependent temperature response curve of  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$ .** The  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  is deposited on a Si substrate with an additional Al top layer (100 nm). The data are obtained by the ultrafast laser flash method (rear-face heating, front-face detection) using nanosecond pulse heating. The signal data are enlarged around the instant of pump-laser irradiation at  $t = 0$ . Inset, the entire signal from one pulse to just before the next pulse. The red solid lines are a least-squares fit.

substantial enhancement of the thermopower takes place if the Heusler system is sputtered on the Si substrate. Recently it has been observed that perovskite-based substrates boost the thermopower of cobaltite thin films<sup>14</sup>, resulting in a >300% increase in the power factor PF of the cobaltite. An enormous value of the figure of merit,  $ZT \approx 2.7$ , was recently reported for cubic  $\text{Ge}_{1-x}\text{Sb}_x\text{Te}$  deposited on a Si wafer<sup>15</sup>.

The electronic part of the thermoelectric performance is obtained from the power factor  $PF = S^2/\rho$ . Putting both experimental quantities together reveals the temperature dependence of PF, plotted in Fig. 2c. A narrow range where PF reaches very large values—more than  $40 \text{ mW m}^{-1} \text{ K}^{-2}$ —is determined to lie between roughly  $50^\circ\text{C}$  and  $150^\circ\text{C}$ . Such values are about ten times larger than those obtained for Bi–Te-based materials<sup>16,17</sup>.

To derive the thermal conductivity of thin-film  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  both diffusivity and effusivity measurements were carried out at room temperature, using an ultrafast laser flash method<sup>18</sup> and a picosecond thermoreflectance measurement method<sup>18,19</sup>, respectively.

A typical transient temperature curve taken at  $T = 300 \text{ K}$  is shown in Fig. 3. The solid red line shows the theoretical evolution of the time-dependent temperature  $T(t)$  of the front face if the rear face is heated<sup>18</sup>. The least-squares fit in Fig. 3 (discussed in detail in Methods), reveals the thermal conductivity as determined from the measured thermal diffusivity to be  $\lambda_{\text{diff}} = 2.70 \text{ W K}^{-1} \text{ m}^{-1}$ . This value is more than 25% smaller than the respective figure derived for bulk  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  (ref. <sup>20</sup>).

Employing a front-heating–front-detection setup, the effusivity  $e$  can be obtained<sup>18,19</sup>, yielding  $\lambda_{\text{eff}} = 3.02 \text{ W K}^{-1} \text{ m}^{-1}$  (where  $\lambda_{\text{eff}}$  is the thermal conductivity determined from the measured thermal effusivity), which is in very good agreement with the value derived from the thermal diffusivity data. Extended Data Fig. 7a–c and Methods summarize the experimental data and contain a thorough discussion.

Assuming, as a first approximation, the above-indicated thermal conductivities at room temperature for  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$ , the approximated figure of merit  $ZT_{\text{appr}}$  can be evaluated. Results are shown in Fig. 2d. The moderate value of  $\lambda$ , in the context of the very large power factor, enables the figure of merit to attain values around or even above 5; such values are well above all others reported so far in the literature. Among the largest values obtained recently are  $ZT = 2.4$  in artificial layers

of  $\text{Bi}_2\text{Te}_3$  and  $\text{Sb}_2\text{Te}_3$  at room temperature<sup>7</sup>,  $ZT = 2.6$  in SnSe along the  $b$  axis of the unit cell<sup>8,9</sup>, and  $ZT = 2.5$  in p-type PbTe–SrTe<sup>10</sup>.

Using the Wiedemann–Franz law, the data for the temperature-dependent  $\lambda_e$  can be assessed and added to  $\lambda_{\text{ph}}$ , which is assumed to be constant. This procedure enables us to derive another set of  $ZT_{\text{appr}}(T)$  values, shown by the open diamonds in Fig. 2d. Again,  $ZT_{\text{appr}}$  reaches values near 5 around  $350\text{--}400 \text{ K}$ ; the range in which Bi–Te-based systems have their best performance.

The small amount of active thin-film material compared to the thickness of the substrate, means care will be required when it is used in thermoelectric applications<sup>21</sup>. Certainly, when integrated in thin-film devices and sensors, using thin-substrate structures might be beneficial.

## Discussion

The thermoelectric performance of the full-Heusler-based thin film deposited on a Si wafer, although unexpected, may have two driving forces. First, the change of crystal structure from fcc in the bulk material to bcc in the thin film. The bcc-type crystal structure, although metastable in the bulk state, becomes stabilized if the material is deposited as a film on the Si substrate. As a consequence, beneficial modifications of the electronic and phononic properties might occur, resulting in the increase in the thermoelectric figure of merit we observe. Second, the electronic structure is advantageous with respect to electronic transport in the material. In this context, the very large values of  $S(T)$  can, at least qualitatively, be understood from the large variation in the electronic DOS near the Fermi energy. Mott’s theory of thermopower, that is

$$S(T) = -\mathcal{A}T \frac{1}{N(E)} \frac{\partial N(E)}{\partial E} \Big|_{E=E_F} \quad (3)$$

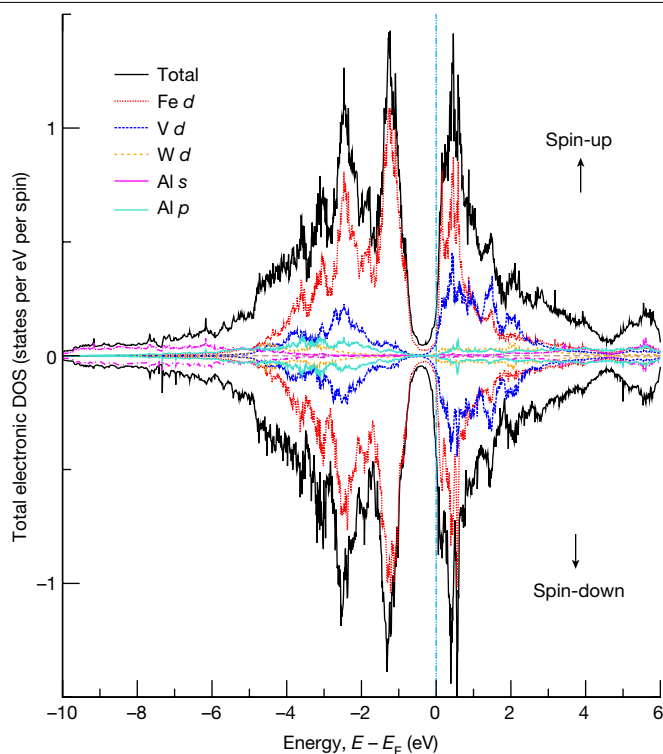
demonstrates that the absolute Seebeck values of a system are proportional to the logarithmic derivative of the density of states  $N(E)$  with respect to energy at the Fermi energy  $E_F$  ( $\mathcal{A}$  is a constant and  $T$  is the temperature). The signature of the thermopower follows here from the signature of  $\partial N(E)/\partial E$ .

To obtain the specific electronic DOS and the respective electronic structure, we performed DFT calculations by constructing 80-atom bcc-type supercells, with an experimental composition of  $\text{Fe}_{0.5}\text{V}_{0.2}\text{W}_{0.05}\text{Al}_{0.25}$ .

Spin-polarized calculations revealed that 25% of the Fe atoms carry small local spin moments of about  $0.2\text{--}0.4\mu_B$  (where  $\mu_B$  is Bohr’s magneton), when doped W atoms are their nearest neighbours, with the shortest Fe–W bonding length about  $2.46\text{--}2.49 \text{ \AA}$ ; the remaining Fe atoms and all the V, W and Al atoms do not have any local spin moment. As shown in Fig. 4, the majority spin-up and minority spin-down densities do not show distinct differences. The electronic states at the Fermi level are dominated by contributions from Fe, V and W; the contributions from Al are almost negligible.

The electronic DOS exhibits an apparent and deep pseudo-gap of about  $0.3\text{--}0.4 \text{ eV}$ , just below the Fermi level. More importantly, at the Fermi level the total DOS is located at an extremely steep shoulder with a very large slope (Fig. 4), that is, at  $\partial N(E)/\partial E|_{E=E_F} = 1.39 \text{ states per eV}^2$  for the spin-up channel and  $4.54 \text{ states per eV}^2$  for the spin-down channel. Following equation (3), this agrees with the experimentally observed very large negative values of the Seebeck coefficient. The narrow region, where very large  $S(T)$  values are obtained, is thought to result from a combination of the large logarithmic derivative of the electronic DOS (equation (3)) and from the fact that owing to the narrowness of the energy gap, both electrons and holes are involved as charge carriers, because electrons are thermally excited across the band gap. This leads to a deviation from linearity as inferred from equation (3) and to a reduction in the absolute values of  $S(T)$  at higher temperatures owing to an increased number of charge carriers. The latter





**Fig. 4 | Electronic DOS of  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$ .** The electronic DOS is derived from the presently obtained lowest-energy 80-atom bcc-type supercell and is shown below and above the Fermi energy (where positive densities represent the majority spin-up channel and negative densities correspond to the minority spin-down channel). The total DOS and the contributions of the various atoms and the partial DOS from which they originate are shown by the different colours. The blue vertical line marks the Fermi energy,  $E_F$ .

is fairly well reflected from a two-carrier analysis of Hall data, explained in detail in Methods.

In order to explain the effect of substituted W on the band structure of  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$ , we built an artificial tetragonal  $\text{Fe}_2\text{VAl}$  parent compound composed of two bcc-type cells (see Extended Data Fig. 9a). Although spin-polarized calculations show that tetragonal  $\text{Fe}_2\text{VAl}$  is non-magnetic, a Dirac-nodal-line-like feature<sup>22,23</sup> appears on its band structure near the Fermi level owing to both a linear band crossing and the negligible spin-orbit coupling. After W doping, we observed two distinct differences in the electronic structure of the 80-atom bcc-type supercell compared to the parent compound. In the first, magnetic moments develop at some Fe sites, as discussed above. In the second, with respect to the case without W doping, the Fermi level is upshifted, resulting in the notably extension of the electron pockets near the  $\Gamma$  and X points (Extended Data Fig. 10a,b). It is mainly because the minima of the energies of the electronic pockets are shifting down with respect to the case without W doping. More importantly, the W-induced magnetic ordering splits the Dirac-nodal-line-like band structure, and as a result Weyl nodes seem to emerge. The Weyl nodes are closer to the Fermi level than are the Dirac nodal lines in the parent material where W is absent. In other words, bcc-type  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  exhibits potential Weyl-like fermions around the Fermi level for both the spin-up and spin-down channels, thereby leading to a possible profound, non-trivial topology of its electronic band structure. The appearance of the nearly linear crossings for both spin channels indicates the enhanced mobilities of charge carriers and constitutes the basics of topological fermions, in analogy with Dirac or Weyl fermions<sup>24,25</sup>, responsible for the remarkable transport properties observed. In such systems, both Weyl-like fermions and its non-trivial surface electronic states are highly robust and are protected from backscattering by non-magnetic disorder and

defects. The scenario outlined here might be related to mechanisms behind the thermoelectric performance of  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$ , as experimentally observed in this study. Theoretical considerations of topological insulators support such conceptions, and ZT values as large as 20 have been proposed<sup>26</sup>.

As demonstrated in several recent studies<sup>27–29</sup> a multi-valley structure of electronic bands near the Fermi energy is beneficial for increased thermopower. Taking into account the DFT results of Extended Data Fig. 10a, b, there are, besides a narrow hole structure, several highly degenerate electronic valleys located near the Fermi energy. The respective charge carriers are characterized by large Fermi velocities and large mobilities (as discussed in detail in Methods). As a result, the dominant diffusion component of the Seebeck effect is also large.

In summary, measurements of the electrical resistivity, thermal conductivity and the Seebeck effect have revealed very large values of the thermoelectric figure of merit for  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  thin films deposited on a Si substrate. Electron microscopy reveals a narrow diffusion zone between the Heusler thin film and the Si substrate; in addition, the ordinary structure of Heusler alloys (fcc,  $\text{Cu}_2\text{MnAl}$ -type) transforms to a bcc W-type structure, which is metastable in the bulk. DFT calculations reveal a large slope of the electronic DOS at nearly the Fermi energy, in agreement with the large Seebeck values observed, as well as Weyl-like fermions similar to topological Weyl semimetals, known for very large charge carrier mobilities. This is a prerequisite for high-performing thermoelectrics.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1751-9>.

- Cahill, D. G., Watson, S. K. & Pohl, R. O. Lower limit to the thermal conductivity of disordered crystals. *Phys. Rev. B* **46**, 6131–6140 (1992).
- Dresselhaus, M. et al. New directions for low-dimensional thermoelectric materials. *Adv. Mater.* **19**, 1043–1053 (2007).
- Koumoto, K. & Mori, T. (eds) *Thermoelectric Nanomaterials* (Springer, 2013).
- Graf, T., Felser, C. & Parkin, S. S. Simple rules for the understanding of Heusler compounds. *Prog. Solid State Chem.* **39**, 1–50 (2011).
- Felser, C. & Hirohata, A. (eds) *Heusler Alloys: Properties, Growth, Applications* (Springer, 2016).
- Knapp, I. et al. Impurity band effects on transport and thermoelectric properties of  $\text{Fe}_{2-x}\text{Ni}_x\text{VAL}$ . *Phys. Rev. B* **96**, 045204 (2017).
- Venkatasubramanian, R., Siivola, E., Colpitts, T. & O'Quinn, B. Thin-film thermoelectric devices with high room-temperature figures of merit. *Nature* **413**, 597–602 (2001).
- Zhao, L.-D. et al. Ultralow thermal conductivity and high thermoelectric figure of merit in  $\text{SnSe}$  crystals. *Nature* **508**, 373–377 (2014).
- Zhang, H. & Talapin, D. V. Thermoelectric tin selenide: the beauty of simplicity. *Angew. Chem.* **53**, 9126–9127 (2014).
- Tan, G. et al. Non-equilibrium processing leads to record high thermoelectric figure of merit in  $\text{PbTe-SrTe}$ . *Nat. Commun.* **7**, 12167 (2016).
- Mori, T. Novel principles and nanostructuring methods for enhanced thermoelectrics. *Small* **13**, 1702013 (2017).
- Chen, Y.-T. & Tan, Y. The optical, magnetic, and electrical characteristics of  $\text{Fe}_2\text{Si}$  thin films. *J. Alloys Compd.* **615**, 946–949 (2014).
- Tang, C. P., Tam, K. V., Xiong, S. J., Cao, J. & Zhang, X. The structure and electronic properties of hexagonal  $\text{Fe}_2\text{Si}$ . *APL Adv.* **6**, 065317 (2016).
- Yordanov, P. et al. Perovskite substrates boost the thermopower of cobaltate thin films at high temperatures. *Appl. Phys. Lett.* **110**, 253101 (2017).
- Wong, D. High power factor Ge-Sb-Te thermoelectric thin film: an evidence of temperature-induced band convergence. In *37th Int. Conference on Thermoelectrics (ICT, 2018)*.
- Cha, J., Zhou, C., Cho, S.-P., Park, S. H. & Chung, I. Ultrahigh power factor and electron mobility in n-type  $\text{Bi}_2\text{Te}_{3-x}\text{Cu}$  stabilized under excess Te condition. *ACS Appl. Mater. Interfaces* **11**, 30999–31008 (2019).
- Hazama, H. et al. Improvement of power factor of n-type  $\text{Bi}_2\text{Te}_3$  by dispersed nanosized  $\text{Ga}_2\text{Te}_3$  precipitates. *J. Alloys Compd.* **726**, 578–586 (2017).
- Baba, T., Taketoshi, N. & Yagi, T. Development of ultrafast laser flash methods for measuring thermophysical properties of thin films and boundary thermal resistances. *Jpn. J. Appl. Phys.* **50**, 11RA01 (2011).
- Baba, T. Analysis of one-dimensional heat diffusion after light pulse heating by the response function method. *Jpn. J. Appl. Phys.* **48**, 05EB04 (2009).



20. Hinterleitner, B. et al. Stoichiometric and off-stoichiometric full Heusler  $\text{Fe}_2\text{V}_{1-x}\text{W}_x\text{Al}$  thermoelectric systems. Preprint at <https://arxiv.org/abs/1801.08966> (2018).
21. Petsagkourakis, I. et al. Thermoelectric materials and applications for energy harvesting power generation. *Sci. Technol. Adv. Mater.* **19**, 836–862 (2018).
22. Burkov, A. A., Hook, M. D. & Balents, L. Topological nodal semimetals. *Phys. Rev. B* **84**, 235126 (2011).
23. Li, R. et al. Dirac node lines in pure alkali earth metals. *Phys. Rev. Lett.* **117**, 096401 (2016).
24. Lv, B. Q. et al. Experimental discovery of Weyl semimetal TaAs. *Phys. Rev. X* **5**, 031013 (2015).
25. Weng, H., Fang, C., Fang, Z., Bernevig, B. A. & Dai, X. Weyl semimetal phase in noncentrosymmetric transition-metal monophosphides. *Phys. Rev. X* **5**, 011029 (2015).
26. Xu, N., Xu, Y. & Zhu, J. Topological insulators for thermoelectrics. *npj Quant. Mater.* **2**, 51 (2017).
27. Tang, Y. et al. Convergence of multi-valley bands as the electronic origin of high thermoelectric performance in  $\text{CoSb}_3$  skutterudites. *Nat. Mater.* **14**, 1223–1228 (2015).
28. Zeier, W. G. et al. Thinking like a chemist: intuition in thermoelectric materials. *Angew. Chem.* **55**, 6826–6841 (2016).
29. Zhang, J. et al. Discovery of high-performance low-cost n-type  $\text{Mg}_3\text{Sb}_2$ -based thermoelectric materials with multi-valley conduction bands. *Nat. Commun.* **8**, 13901 (2017).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

## Methods

### Experimental details

Thin film samples were prepared by magnetron sputtering using a single target consisting of the stoichiometrically prepared bulk sample in the form of a disk 25 mm in diameter and 3 mm in height. The base pressure was  $10^{-4}$  Pa, the working gas was Ar with a pressure of 2 Pa and the distance from target to substrate was about 3 cm. An undoped silicon wafer with a thickness of 0.279 mm and [100] orientation was used as substrate (Siebert Wafer, [https://www.siebertwafer.com/Silicon\\_Wafers.html](https://www.siebertwafer.com/Silicon_Wafers.html)). The electrical resistivity of the wafer was  $>100 \Omega \text{ cm}$  at room temperature. The substrate temperatures were varied up to 650 °C and the Heusler alloy was deposited on the substrate with a deposition rate of  $0.5 \text{ nm s}^{-1}$  to create samples with thicknesses ranging from a few hundred nm to 3  $\mu\text{m}$ . To stabilize the crystal structure and recrystallize the amorphous state, the samples were heat treated for one week at temperatures ranging from 150 °C to 650 °C.

The phase purity of the samples and the lattice parameter were verified by X-ray diffraction, using Cu K $\alpha$  radiation (D5000, Siemens) and electron diffraction. Electron microscopy was used to obtain chemical information at the film–substrate interface. For this, a field emission gun transmission electron microscope (FEI TECNAI F20) was used. For chemical analysis an electron filter was attached (GATAN GIF Tridiem). The samples were mechanically thinned to a thickness of 10  $\mu\text{m}$  and further processed within an ion mill (GATAN PIPS). The last polishing step used 200-eV Ar $^{+}$  ions to remove beam damage from the prior milling process.

Transport data (the resistivity and Seebeck effect above room temperature) were taken using an ULVAC ZEM-3. The Hall effect above room temperature was studied using a home-made set-up based on a superconducting magnet (up to 12 T). The electrical resistance and the Hall resistance were derived using the van der Pauw technique, using an a.c. resistance bridge (LakeShore 370).

Thermal diffusivity was measured at room temperature using the ultrafast laser flash method (rear heating–front detection)<sup>18</sup>. The thermoelectric film is heated by a nanosecond laser pulse through the Si substrate, which is transparent at the wavelength of the heating laser beam. Opposite to the heating beam, a time-dependent temperature response is detected via the thermoreflectance technique by a probing laser beam, which is illuminated onto the Al surface (thickness, 100 nm) that is deposited on the thermoelectric film. The ultrafast laser flash method is a natural extension of the laser flash method, which is a standard method used to measure thermal diffusivity of bulk materials. This method is established as one of the metrological standards of thermal diffusivity of thin films under metric convention maintained by the Bureau International des Poids et Mesures (BIPM). It is considered to be the standard method to measure thermal diffusivity of films as thin as several hundred nanometres<sup>30</sup>. A reference sample for the ultrafast laser flash method—680-nm TiN on a synthetic quartz substrate—was previously developed as a thin-film thermal conductivity standard sample, and supplied by the National Metrology Institute of Japan<sup>31</sup> (see Extended Data Fig. 8).

The thermal effusivity was obtained by applying the picosecond time-domain thermoreflectance technique using a NanoTR (PicoTherm) and a customized thermal analysis system based on PicoTR (PicoTherm). The customized system enables a selective derivation of the thermal effusivity of the thin-film area by focusing the probe laser beam. Thus, one can even detect time-domain thermoreflectance signals from sample surfaces that are somewhat rough and that have only a narrow terrace structure. The 100-nm Al thin film deposited on the sample surface acts as a heat source of known areal heat capacity.

Specific heat data on bulk  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  were collected from a differential scanning calorimetry measurement (Linseis, DSC-PT10).

### Ab initio calculations

Within the framework of DFT<sup>32,33</sup>, we performed calculations for structural optimization and the electronic band structures. DFT calculations were performed using the Vienna ab initio Simulation Package (VASP)<sup>34–36</sup>, with projector augmented wave pseudo-potentials<sup>37,38</sup> and the generalized gradient approximation within the Perdew–Burke–Ernzerhof exchange–correlation functional<sup>39</sup>. The adopted pseudo-potentials of all elements treat semi-core valence electrons as valence electrons. An accurate optimization of structural parameters was calculated by minimizing the interionic forces below  $0.0001 \text{ eV } \text{\AA}^{-1}$ . The cut-off energy for the expansion of the wave function into the plane waves was 400 eV. The Brillouin zone integrations were sampled with a resolution of  $2\pi \times 0.014 \text{ \AA}^{-1}$ . To analyse the effects of W doping on the band structure of the supercell, we adopted the unfolding technique implemented in VASP<sup>40,41</sup>.

Theoretical calculations of  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  within the random bcc-type structure are difficult: using currently available theoretical techniques for DFT, it is impossible to consider the random elemental distribution in  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  with sufficient accuracy; that is, although the periodic unit cells have finite lattice sites, there are too many possible atomic distributions of Fe, V, W and Al to sample them all. Therefore, in order to simulate the situation as close as possible to reality, we first constructed a  $3 \times 3 \times 3$  54-atom bcc supercell to investigate the lowest-energy configuration by considering the atomic distribution and nearest-neighbours among Fe, V, W and Al. The results suggest that the lowest-energy 54-atom configuration is one in which W tends to bind with nearest-neighbour V atoms in the same atomic layer and where the W–V layer is in between two atomic Fe layers such that each V and W atom has at least one neighbouring Fe atom. All Al atoms occupy the same atomic layers, located between two atomic Fe layers. Based on this lowest-energy 54-atom configuration, and in order to consider more possible configurations, we additionally constructed an 80-atom bcc-type supercell. Within this 80-atom supercell, we varied the Al, V and W positions in the bcc-type Fe framework—Fe is known to have an inherent stable bcc ground state. In total, we constructed 45 different configurations; from these we found an energetically favourable 80-atom supercell with 40 Fe, 16 V, 4 W and 20 Al atoms, a good match to the experimental composition of  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$ .

**The film–substrate interface.** The deposition of the Heusler film  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  (thickness, 1  $\mu\text{m}$ ) on top of a cleaned (100) Si substrate (thickness, 279  $\mu\text{m}$ ) forms a composite (see Fig. 1e, f), consisting of the active thermoelectric layer, the interface and the substrate. The interface is created by diffusion processes in both directions (that is, from the layer to the substrate and vice versa), because of elevated temperatures during the sputtering process and during the heat treatment. The emergence of products resulting from diffusion in both directions follow the general thermodynamic rules of phase formation. Respective phase diagrams<sup>42</sup> indicate that Al and Si form a eutectic system at 12.6 wt% Si at 577 °C. In addition, Si does not solvate any Al. As a consequence, Al should not diffuse into the Si substrate. On the other hand, Fe–Si, Si–V and Si–W constitute several binary compounds; details on the various phases can be found, for example, in ref. <sup>42</sup>.

Electron-microscopy-based studies we have carried out on the thin Heusler film deposited on Si indicate that there are two main interface structures established between Si and the Heusler film: an  $\text{Al}_2\text{O}_3$  interlayer and binary  $\text{Fe}_3\text{Si}$ . While the former has a thickness of about 10 nm, the latter extends about 20 nm into the Si substrate (see Fig. 1e).

ELNES of the interfacial atomic layers enables us to understand the chemical bonding at the interface. Extended Data Fig. 1a illustrates the oxygen K-edge and the vanadium L-edge at 532 eV and 508 eV energy loss, respectively. Extended Data Fig. 1b shows the iron L-edge at 708 eV energy loss for both  $\text{Fe}_2\text{Si}$  and the Heusler thin film. The corresponding spectra were used to fit the phase distribution in Fig. 1f.

**Thermoelectric properties.** Studies of the electronic and thermal transport were performed on samples with about 10 mm length and 4 mm width. A scheme of such a measurement is shown in Extended Data Fig. 2. This schematic drawing demonstrates that both the heat and the electric current is flowing from the upper to the lower part of the sample, constituting a parallel arrangement of three different layers: the active thermoelectric material, the interface layer and the Si wafer. In order to derive individual contributions, the electrical resistivities can be considered in terms of classical parallel circuits. However, the Seebeck effect requires special consideration<sup>43</sup>. Accordingly, the Seebeck effect obtained consists of each individual contribution weighted by the respective electrical conductivity  $\sigma$  and the thickness  $d$  of the layers.

To isolate the contribution associated with the Si substrate, both the Heusler film and the interface layer were removed mechanically by grinding; the substrate was thinned by about 5  $\mu\text{m}$ . Transport data derived for the Si substrate are summarized in Extended Data Fig. 3a, b.

The procedure to analyse the thermopower data of thin films deposited on substrates in terms of equation (2) has been applied recently to various systems, such as  $\text{Ca}_3\text{Co}_4\text{O}_9$  on  $\text{SrTiO}_3$  or  $\text{LaAlO}_3$ <sup>14</sup>, and used to determine the effect of oxygen reduction in bulk  $\text{SrTiO}_3$  substrates when a thin film was deposited in an oxygen-deficient environment<sup>44</sup>, in Si–Ge superlattices<sup>45</sup>, and in high-temperature superconducting films<sup>46</sup>. It is based on the assumptions behind the so-called Kohler rule, which develops into the Nordheim–Gorter rule. The Nordheim–Gorter rule is the analogue for thermopower of Matthiessen’s rule for electrical resistivity.

There is, on the one hand, a substantial difference (several orders of magnitude) between the temperature-dependent resistivities of the composite of Heusler film, interface and Si substrate compared to the Si substrate alone. As a result, charge carriers predominantly pass through the Heusler film, dominating the electrical conductivity. On the other hand, such distinct differences are absent in the Seebeck effect. As is clear from equation (2), separating the individual thermopower contributions requires determining the respective electrical conductivities,  $\sigma = 1/\rho$ , and the respective layer thicknesses. Equation (2) shows that from the huge differences in conductivities (more than five orders of magnitude in the lower ranges of temperature), the contribution of the Si substrate to the overall measured effect of the composite remains small, even considering that the substrate is about a hundred times thicker than the Heusler layer. A similar argument holds for the interface and its contribution to the measured data: even assuming very large Seebeck values for metallic  $\text{Fe}_2\text{Si}$  (for example,  $100 \mu\text{V K}^{-1}$ ) and small resistivities (for example,  $100 \mu\Omega \text{cm}$ ), the two-orders-of-magnitude-smaller thickness of this interface does not overwhelmingly contribute to the observed Seebeck data. As a result, the deduced  $S(T)$  values of the Heusler film are similar to those obtained from measurements of the entire set-up.

If the contribution of the interlayer is neglected—the island-like structures are only weakly connected to each other—only small changes in  $S(T)$  would result, as demonstrated in the main text.

The simple overall behaviour of  $S(T)$  as described by Mott’s equation (equation (3)) is modified by the fact that the observed thermopower does not only consist of contributions from a single charge carrier type. Instead, owing to the relatively narrow gap in  $N(E)$  next to the Fermi energy, both electrons and holes become actively involved, that is,

$$S = \frac{\sum_i \sigma_i S_i}{\sum_i \sigma_i} \quad (4)$$

Here,  $\sigma_i$  and  $S_i$  are the electrical conductivities and the Seebeck coefficients, respectively, of the different sets of charge carriers.

Equation (4) indicates that the measured thermopower data of narrow-gap semiconductors might exhibit complicated behaviours—the

almost linear temperature dependence, as inferred from equation (3), can be very different, owing to the temperature-dependent change in the charge carrier density  $n$  as well as owing to the temperature-dependent electron–phonon interaction that influences the electrical conductivity of a system.

**High-temperature and high-field Hall studies.** The following subsections are based on the data obtained from additionally prepared thin films based on  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  deposited on a Si substrate and heat-treated at  $450^\circ\text{C}$ .

To derive qualitative and quantitative information about the dominant charge carriers present in these thin-film Heusler systems, we carried out Hall effect measurements at temperatures above room temperature and at magnetic fields up to 11 T, employing the van der Pauw technique; results are shown in Extended Data Fig. 4a. Besides a small positive contribution for the room temperature run, all remaining Hall resistivities,  $\rho_{xy}(B)$ , are negative; however, they exhibit strong curvatures with increasing magnetic field. (Here,  $\rho_{xy}(B)$  characterises an electrical resistivity measured along the  $y$  direction when an electrical current is flowing along the  $x$  direction and the magnetic field  $B$  is in the  $z$  direction.) This behaviour indicates that there is no unique set of charge carriers—instead, different types are present, for example, electrons and holes, or a set of electrons with different mobilities.

To account theoretically for  $\rho_{xy}(B)$ , a simplified two-carrier model<sup>47,48</sup> is generally used, with four adjustable quantities. An improvement of this model<sup>49</sup> has reduced the number of fit parameters to two, and two experimentally obtainable quantities serve as input parameters.

The solid lines in Extended Data Fig. 4a are least-squares fits to the data, which gives the temperature-dependent majority and minority charge carrier densities  $n_1$  and  $n_2$  and the (respective) mobilities  $\mu_1$  and  $\mu_2$ , as summarized in Extended Data Fig. 4b. For the high-temperature range, there are two different sets of electrons: the set with higher charge carrier concentration exhibits lower mobilities, and the set with lower charge carrier concentration is characterized by very high mobility. The temperature dependence of  $n_1$  reveals a weak minimum around  $100^\circ\text{C}$ , roughly corresponding to the extremum of the temperature-dependent Seebeck coefficient. Within the two-carrier model, at room temperature the positive low-field data requires a small concentration of holes with very high mobility as charge carriers ( $n_2 = 1.03 \times 10^{18} \text{cm}^{-3}$  and  $\mu_2 = 5,920 \text{cm}^2 \text{V}^{-1} \text{s}^{-1}$ ) and a much larger content of electrons with reduced mobilities ( $n_1 = 1.1 \times 10^{22} \text{cm}^{-3}$  and  $\mu_1 = -7.5 \text{cm}^2 \text{V}^{-1} \text{s}^{-1}$ ).

To compare the electronic transport in both film and bulk on a more microscopic scale, Hall measurements of bulk  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  were carried out (the results are summarized in Extended Data Fig. 5). Notably,  $\rho_{xy}(B)$  is negative and almost linear for the given field and temperature range, indicating that electrons are the dominant charge carriers. Using a standard analysis for the field-dependent Hall data yields a charge carrier density at  $T = 300 \text{K}$  of  $n_{\text{bulk}} = 5.18 \times 10^{21} \text{cm}^{-3}$  and a mobility of  $\mu_{\text{bulk}} = -3.29 \text{cm}^2 \text{V}^{-1} \text{s}^{-1}$ . At  $T = 380 \text{K}$ , the following values are found:  $n_{\text{bulk}} = 7.6 \times 10^{21} \text{cm}^{-3}$  and  $\mu_{\text{bulk}} = -2.13 \text{cm}^2 \text{V}^{-1} \text{s}^{-1}$ . The same analysis for the film sample with  $B \rightarrow 0$  and at  $T = 373 \text{K}$  gives  $n_{\text{film}} = 1.95 \times 10^{20} \text{cm}^{-3}$  and  $\mu_{\text{film}} = -90 \text{cm}^2 \text{V}^{-1} \text{s}^{-1}$ ; and at  $T = 469 \text{K}$ ,  $n_{\text{film}} = 3 \times 10^{20} \text{cm}^{-3}$  and  $\mu_{\text{film}} = -67 \text{cm}^2 \text{V}^{-1} \text{s}^{-1}$ . These numbers indicate that the charge carrier density in the bulk material is about one order of magnitude larger than in the film, but the mobility is substantially smaller.

Simple arguments enable us to exclude the Si substrate as being responsible for the considerable differences between the film and the bulk data (Extended Data Figs. 4, 5). Assuming two dominant sets of charge carriers, one for the film (A) and one for the substrate (B), the observed Hall coefficient  $R_H$  can be expressed as<sup>43</sup>

$$R_H = \frac{\frac{R_{HA}}{\rho_{xxA}^2} + \frac{R_{HB}}{\rho_{xxB}^2}}{\left(\frac{1}{\rho_{xxA}} + \frac{1}{\rho_{xxB}}\right)^2} \quad (5)$$

Taking into account the electrical resistivity of the Si wafer,  $\rho_{\text{SiB}} > 10^8 \mu\Omega \text{ cm}$ , compared to the Heusler film resistivity (several hundred  $\mu\Omega \text{ cm}$ ), it follows that the second term in the numerator of equation (5) tends towards zero, although the mobility of Si is two or three orders of magnitude larger, and the second term in denominator is negligible compared to the first. Altogether, this comparison implies that the Si substrate provides only a minor contribution to the Hall data, and thus  $R_{\text{H}}$  as obtained from the measurement is dominated only by the Heusler film.

**Stability and reliability of measurements and stability of the samples.** The accuracy of electronic and thermal transport can suffer from contact problems, and so a series of Seebeck and electrical resistivity measurements was carried out several times, increasing the temperature to above 400 °C and then cooling it down to room temperature. This procedure enables us to confirm the measurement reliability and the chemical and thermal stability of the thin films.

The results of the Seebeck coefficient measurements from these experiments are plotted in Extended Data Fig. 6a. The overall behaviour of each individual run matches the results of the other runs; although there are some changes in absolute values, the variance remains small. These runs demonstrate that in the given temperature range the behaviour of the material is stable, although the distinct bcc structure is stable only in conjunction with the Si substrate. In addition, the various runs show that there is no diffusion of the Heusler phase into the substrate (or vice versa), which would affect electronic transport in the films. Furthermore, these measurements reveal a very similar temperature dependence of  $S(T)$  compared to the main results we report here. We thus also demonstrate that thin films with similar quality can be repeatedly prepared in the set-up at TU Wien.

Extended Data Fig. 6b displays the temperature-dependent electrical resistivity of the same sample taken during the Seebeck effect experiments. Again, the overall behaviour of  $\rho(T)$  remains almost unchanged; although slightly larger variations than seen in  $S(T)$  are present, which might indicate slight changes in the mechanical quality of the sample. Extended Data Fig. 6c represents the power factor  $\text{PF} = S^2/\rho$  of the same sample as evaluated from the data of Extended Data Fig. 6a, b. Similar to the data reported in the main text, the power factor is very high and therefore will also give very high values of  $ZT$ .

Data presented in this section verify the chemical and thermal stability of the thin films based on  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$ . Furthermore, the data demonstrate the repeatability of the preparation technique we use to deposit the thin film and the soundness of our data acquisition.

**Thin-film thermal conductivity.** To derive the thermal conductivity of thin film  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$ , the ultrafast laser flash method and a picosecond thermoreflectance study were carried out at room temperature. A typical result of the flash method is displayed in Fig. 3.

The least-squares fit in Fig. 3 is derived by adjusting two parameters, the heat diffusion time  $\tau_{\text{diffusion}} = d^2/\alpha$  across the specimen and the cooling time constant for effusion to the Si substrate  $\tau_{\text{cooling}}$ , where  $d$  is the thickness of the sample and  $\alpha$  is the thermal diffusivity. The heat diffusion time, from the rear face of the thermoelectric film to the Al thin-film surface, was obtained from this fit as  $\tau_{\text{diffusion}} = 1.44 \times 10^{-6} \text{ s}$ , with a standard deviation of  $1.09 \times 10^{-7} \text{ s}$ . The cooling time constant is determined to be  $\tau_{\text{cooling}} = 3.63 \times 10^{-6} \text{ s}$  with a standard deviation of  $0.36 \times 10^{-6} \text{ s}$ . Using the thermal diffusivity of Al<sup>50</sup> ( $\alpha = 9.7 \times 10^{-5} \text{ m}^2 \text{ s}^{-1}$ ), the heat diffusion time is  $\tau_{\text{Al}} = d^2/\alpha = 1.03 \times 10^{-10} \text{ s}$  for  $d_{\text{Al}} = 100 \text{ nm}$ . This indicates that  $\tau_{\text{Al}} \ll \tau_{\text{specimen}}$ . Hence, the Al layer on top of the Heusler film does not substantially affect the present analysis and  $\alpha_{\text{specimen}} \approx \alpha_{\text{Heusler}} = (8.49 \pm 0.739) \times 10^{-7} \text{ m}^2 \text{ s}^{-1}$ .

The thermal conductivity is then

$$\lambda_{\text{diff}} = \alpha DC \quad (6)$$

where  $D$  is the density and  $C$  the volumetric heat capacity of the system. Using the theoretical density of  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$ ,  $D = 7,406 \text{ kg m}^{-3}$ ,

and the heat capacity (derived experimentally),  $C = 430 \text{ J kg}^{-1} \text{ K}^{-1}$ , gives  $\lambda_{\text{diff}} = 2.70 \text{ W K}^{-1} \text{ m}^{-1}$ . This value is more than 25% smaller than the figure derived for bulk  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  (ref. <sup>20</sup>).

Thermal conductivity values of the order of  $1.5\text{--}3 \text{ W K}^{-1} \text{ m}^{-1}$  have been previously reported<sup>51</sup> for  $\text{Fe}_2\text{VAl}$  deposited as thin film on either  $\text{MgAl}_2\text{O}_4$  or  $\text{MgO}_2$ . In that work, the more than tenfold reduction in  $\lambda(T)$  of thin film  $\text{Fe}_2\text{VAl}$  compared to bulk  $\text{Fe}_2\text{VAl}$  was attributed to structure disordering, including dislocations, atomic vacancies, stacking faults, and other factors introduced during the preparation of the film<sup>51</sup>. As inferred from our X-ray results, there is not only disorder due to the V or W substitutions, but additionally all atoms in the unit cell are randomly distributed at the same (2a) bcc-type lattice site. As a consequence, the heat-carrying phonons are expected to be maximally scattered. Thus, the lattice thermal conductivity should be lower than in the bulk material. Disorder on all lattice sites in bcc or fcc systems, that is, random solid solutions of elements, is a guiding design concept in high-entropy alloys. These have been proven for very low thermal conductivities<sup>52</sup>. Finally, we note that a substantial reduction of  $\lambda(T)$ —by 1 to 2 orders of magnitude—was found in undoped Si films (thickness,  $\sim 3 \mu\text{m}$ ), compared to bulk Si (ref. <sup>53</sup>).

Extended Data Fig. 7a shows a normalized thermoreflectance signal (front-heating–front-detection setup) of  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$ , enabling us to deduce the effusivity  $\epsilon$ . A 100-nm thick Al surface layer is deposited by magnetron sputtering. The signal has been normalized by subtracting the background offset component. Just after pulse laser irradiation, the surface temperature of the deposited Al film rapidly increases. Then, the surface temperature gradually decreases owing to heat diffusion in the cross-plane direction.

The entire signal, over one cycle from the heating pulse to the next pulse (50 ns), is shown in Extended Data Fig. 7b, c, fitted with a previously reported analytical equation<sup>54,55</sup>. This analytical method considers the thermal effect of the past pulses, which enables us to observe the single-pulse heating effect (shown by the red dotted lines in Extended Data Fig. 7b, c).

The time-dependent thermoreflectance signal contains information about the Heusler film, the Al top layer and has a contribution from the boundary thermal resistance,  $W_{\text{b}}$ .

Typically measured cross-plane thermal conductivity values turn out to be frequently erroneously low, owing to the contribution of  $W_{\text{b}}$ . In general,  $W_{\text{b}}$  is considerably lower than  $1 \times 10^{-7} \text{ m}^2 \text{ K W}^{-1}$ . As a consequence this contribution has to be taken into account when analysing heat transfer across a number of thin films. Following previous work<sup>18,19</sup>, an analytical solution enables us to separate the boundary thermal resistance between the Heusler film and the Al top layer from the thermal effusivity of the thermoelectric film<sup>18</sup>. A least-squares fit gives  $W_{\text{b}} = 1.6 \times 10^{-8} \text{ m}^2 \text{ K W}^{-1}$  and  $\epsilon = 3,100 \text{ J m}^{-2} \text{ s}^{-0.5} \text{ K}^{-1}$  (refs. <sup>54,55</sup>).

The effusivity  $\epsilon$  is related to the thermal conductivity  $\lambda$  by

$$\epsilon = \sqrt{DC\lambda} \quad (7)$$

From equation (7), the thermal conductivity at room temperature is  $\lambda_{\text{eff}} = 3.02 \text{ W K}^{-1} \text{ m}^{-1}$ , in good agreement with the value derived from the thermal diffusivity data. Here, the density of  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$ ,  $D = 7,406 \text{ kg m}^{-3}$ , and the experimentally derived heat capacity,  $C = 430 \text{ J kg}^{-1} \text{ K}^{-1}$ , are used.

Extended Data Fig. 8 shows the signal for the TiN reference-standard sample from the National Metrology Institute of Japan, measured by the ultrafast laser flash system used in this study. A heat diffusion time of  $1.42 \times 10^{-7} \text{ s}$  was obtained by the same method applied to the  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  thin film. This value shows agreement to within 1.6% with the reference value<sup>31</sup>  $\tau_{\text{TiN}} = 1.397 \times 10^{-7} \text{ s}$ . Hence, the thermal diffusivity measurements in this study are comparable to the metrological standard to 2%.

**DFT results.** Because experimental information on specific atomic distributions is still lacking, we considered 45 configurations of atomic



occupations in the 80-atom supercell to obtain the one with the lowest energy. We emphasize that it is impossible to consider all possible configurations, owing to the random occupation of the four elemental constituents (Fe, V, W and Al atoms).

To obtain the doping effects of W with respect to the 80-atom lowest-energy bcc-type supercell with composition  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  (Extended Data Fig. 9b), we adopted the band-unfolding technique to plot the electronic band structure on the basis of the parent unit cell. In this case, we selected bcc-type  $\text{Fe}_2\text{VAl}$  as the parent material. A bcc-type unit cell allows only two atoms, one at the corner and the other at the body centre. Because  $\text{Fe}_2\text{VAl}$  has three types of atoms, we built an artificial tetragonal unit cell (shown in Extended Data Fig. 9a) to use in this process. Note that this tetragonal unit cell can be viewed as two bcc-type unit cells. We determined that this artificial tetragonal  $\text{Fe}_2\text{VAl}$  is non-magnetic, and its electronic band structure is shown by the black curves in Extended Data Fig. 10a, b. Notably, the electronic band structure of this tetragonal  $\text{Fe}_2\text{VAl}$  shows linear band crossings, marked by circles along different high-symmetry lines. The atomic masses of Fe, V and Al are only moderately heavy; only very slight spin–orbit coupling effects are present. Although this artificial  $\text{Fe}_2\text{VAl}$  is non-centrosymmetric, the linear band crossings result in the appearance of so-called topological Dirac nodal lines; on the  $k_y$  plane, that is, on the  $k_z = 0$  plane, perpendicular to the  $k_z$  direction of the Brillouin zone of the tetragonal unit cell, as defined in Extended Data Fig. 9a, there exist four Dirac nodal lines.

After W is used to dope bcc-type  $\text{Fe}_2\text{VAl}$ , the lowest-energy 80-atom supercell was determined from our current first-principles calculations. The effects of W doping are twofold: local magnetic ordering of some Fe atoms occurs when they have nearest neighbouring W atoms and the total electronic valence number increases; it is higher in W than in V, Fe or Al. This increase leads to the upshifting of the Fermi energy. This is well reflected by the electronic band structures of the 80-atom supercell obtained by the band-unfolding technique: in Extended Data Fig. 10a, b we see that around the Fermi level the electronic pockets shift to lower energies for both spin channels as compared to the band structure of the parent material (black curves); indicating the effect of W doping. The occurrence of magnetism from W doping induces the most important effect: the appearance of Weyl nodes around the Fermi level. The appearance of these nodes indicates that spin ordering breaks time-reversal symmetry, which splits the Dirac nodal lines—as seen in the parent  $\text{Fe}_2\text{VAl}$ —into Weyl nodes. In addition, W doping displaces the Weyl nodes closer to the Fermi level, as shown in Extended Data Fig. 10a, b.

From Extended Data Fig. 10a, b we see that the DFT-derived unfolded electronic band structures for the 80-atom bcc supercell of  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  demonstrates the coexistence of the expected multi-valley electron and hole states. Interestingly, there exists a hole pocket with a degeneracy of eight at 0.23 eV above the Fermi level along R–A in the Brillouin zone. In addition, there exist two electronic valleys observed at 0.02 eV and 0.07 eV above the Fermi level, along Z–R and  $\Gamma$ –R, respectively. In particular, the electronic valley along Z–R has a valley degeneracy of eight, whereas the electronic valley along  $\Gamma$ –R path exhibits a valley degeneracy of 16. To visualize the valley pockets along Z–R, R–A and  $\Gamma$ –R, we have compiled their charges of the Fermi surfaces at energies 0.11 eV, 0.11 eV, and 0.13 eV above the Fermi level, respectively (Extended Data Fig. 10c). These distinct valence and conduction band valleys deviate from the high-symmetry points, leading to high valley degeneracies; thus we expect the thermoelectric performance will be enhanced. The carrier mobilities of every type of hole and electronic valley were evaluated from Extended Data Fig. 10a, b. Importantly, at the A point the hole band exhibits an extremely large theoretical mobility ( $1.051 \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1}$  at 300 K) and the other four electronic valleys also have very high electronic mobilities (X,  $0.086 \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1}$ ; M,  $0.545 \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1}$ ; G,  $0.041 \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1}$ ; R,  $0.008 \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1}$ , all at 300 K). The combination of high degeneracies and high mobilities (as obtained from our DFT-derived band structures), and the multi-valley band structure potentially

contributes to a substantial thermoelectric enhancement for bcc-type  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$ , in line with recent reports on multi-valley thermoelectric systems<sup>27–29</sup>. From an electronic structure point of view, there is only a small fraction of holes contributing to the transport properties, and so most contributions are expected to originate from electronic valleys at high temperatures.

On the basis of the Goldsmid formula<sup>56</sup>, the gap in the DOS at the Fermi energy can be estimated as roughly 260 meV, which is in the range calculated by DFT. In addition, the calculations also reveal that the total density at the Fermi level,  $N(E_F) = 0.62$  states per eV per atom within the 80-atom supercell. This would correspond to a Sommerfeld value of the specific heat,  $\gamma_{\text{DFT}} = 1.46 \text{ mJ mol}^{-1} \text{ K}^{-2}$ , in good agreement with experimental data derived for  $\text{Fe}_2\text{VAl}$  ( $\gamma_{\text{expt}} = 1.5 \text{ mJ mol}^{-1} \text{ K}^{-2}$ )<sup>57</sup>.

A substantial enhancement of the Seebeck effect and thus of the thermoelectric performance in the Heusler film studied here, as being as a result of recently discovered<sup>58</sup> phonon drag term in thin-film  $\text{Bi}_2\text{Te}_3$  is unlikely. That work demonstrated that the substrate can play an important role, guiding the phonon drag mechanism in the active film. However, the most prominent enhancement in  $\text{Bi}_2\text{Te}_3$  is observed well below the corresponding Debye temperature  $\theta_D$  ( $\approx \theta_D/10$ ). The thermopower maximum of the  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  thin film at around 100 °C is thus unlikely to be caused by the phonon drag mechanism. Another recently noted mechanism to improve thermoelectricity is superparamagnetism<sup>59</sup> of nanosized particles dispersed in a host material.  $\text{Fe}_2\text{Si}$ , which constitutes the interface layer between the Si substrate and the Heusler film, is a magnetically ordered compound with a Curie temperature of the order of 520 K<sup>60</sup>. However, because  $\text{Fe}_2\text{Si}$  does not form a continuous layer (it instead forms a string-like arrangement of weakly coupled nanosized islands), superparamagnetism might be present in this interface. Our electron microscopy study (see Fig. 2d, e), on the other hand, demonstrates that the  $\text{Fe}_2\text{Si}$  particles are localized and do not diffuse into the Heusler film. Consequently, the superparamagnetism enhancement mechanism is presumably absent in the  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  thin film.

## Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## Code availability

The computer codes that support the findings of this study are available from the corresponding author upon reasonable request.

- Baba, T. et al. Research and development of metrological standards for thermophysical properties of solids in the National Metrology Institute of Japan. *High Temp. High Press.* **39**, 279–306 (2010).
- Yagi, T., Taketoshi, N. & Baba, T. Development of thin film reference material for thermal diffusivity. In *Proc. 1st Int. Symposium on Thermal Design and Thermophysical Property for Electronics* (NMIJ/AIST, 2008).
- Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136**, B864–B871 (1964).
- Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133–A1138 (1965).
- Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47**, 558–561 (1993).
- Kresse, G. & Hafner, J. Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium. *Phys. Rev. B* **49**, 14251–14269 (1994).
- Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
- Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
- Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).
- Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
- Eckhardt, C., Hummer, K. & Kresse, G. Indirect-to-direct gap transition in strained and unstrained  $\text{Sn}_x\text{Ge}_{1-x}$  alloys. *Phys. Rev. B* **89**, 165201 (2014).
- Liu, P. et al. Electron and hole doping in the relativistic mott insulator  $\text{Sr}_2\text{IrO}_6$ : a first-principles study using band unfolding technique. *Phys. Rev. B* **94**, 195145 (2016).
- Massalski, T. B. *Binary Alloy Phase Diagrams* (ASM International, 1990).
- Dugdale, J. *The Electrical Properties of Metals and Alloys* (Dover, 2016).

44. Yu, C., Scullin, M. L., Huijben, M., Ramesh, R. & Majumdar, A. Thermal conductivity reduction in oxygen-deficient strontium titanates. *Appl. Phys. Lett.* **92**, 191911 (2008).
45. Koga, T., Cronin, S. B., Dresselhaus, M. S., Liu, J. L. & Wang, K. L. Experimental proof-of-principle investigation of enhanced  $Z_{30}T$  in (001) oriented Si/Ge superlattices. *Appl. Phys. Lett.* **77**, 1490–1492 (2000).
46. Heinze, S. et al. Thermoelectric properties of  $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$ - $\text{La}_{2/3}\text{Ca}_{1/3}\text{MnO}_3$  superlattices. *Appl. Phys. Lett.* **101**, 131603 (2012).
47. Chambers, R. G. The two-band effect in conduction. *Proc. Phys. Soc. A* **65**, 903–910 (1952).
48. Arushanov, E. K. & Chuiko, G. P. The magnetic field dependence of kinetic coefficients of cadmium arsenide single crystals. *Phys. Status Solidi A* **17**, K135–K138 (1973).
49. Eguchi, G. & Paschen, S. Robust scheme for magnetotransport analysis in topological insulators. *Phys. Rev. B* **99**, 165128 (2019).
50. Jensen, J. E., Tuttle, W. A., Stewart, R. B., Brechna, H. & Prodel, A. G. (eds) *Brookhaven National Laboratory Selected Cryogenic Data Notebook*. Report BNL-10200-R (Brookhaven National Laboratory, 1980).
51. Furuta, Y., Kato, K., Miyawaki, T., Asano, H. & Takeuchi, T.  $\text{Fe}_2\text{VAl}$ -based thermoelectric thin films prepared by a sputtering technique. *J. Electron. Mater.* **43**, 2157–2164 (2014).
52. Tsai, M.-H. Physical properties of high entropy alloys. *Entropy* **15**, 5338–5345 (2013).
53. Cahill, D. G. et al. Nanoscale thermal transport. *J. Appl. Phys.* **93**, 793–818 (2003).
54. Baba, T., Ishikawa, K. & Baba, T. Analysis of heat diffusion in thin films and boundary resistance by pulsed light heating thermoreflectance method. In *3rd Int. Conference on Functional Integrated NanoSystems (NanoFis)*, 2017).
55. Baba, T., Ishikawa, K. & Baba, T. Measurement and analysis of thermal conductivity, thermal diffusivity and interfacial thermal resistance of thermoelectric thin films. In *37th Int. Conference on Thermoelectrics (ICT)*, 2018).
56. Goldsmid, H. J. & Sharp, J. W. Estimation of the thermal band gap of a semiconductor from Seebeck measurements. *J. Electron. Mater.* **28**, 869–872 (1999).
57. Lue, C. S., Ross, J. H., Chang, C. F. & Yang, H. D. Field-dependent specific heat in  $\text{Fe}_2\text{VAl}$  and the question of possible 3d heavy fermion behavior. *Phys. Rev. B* **60**, R13941–R13945 (1999).
58. Wang, G., Endicott, L. & Chi, H. Lošták, P. & Uher, C. Tuning the temperature domain of phonon drag in thin films by the choice of substrate. *Phys. Rev. Lett.* **111**, 046803 (2013).
59. Zhao, W. et al. Superparamagnetic enhancement of thermoelectric performance. *Nature* **549**, 247–251 (2017).
60. Varga, L. K., Mazaleyrat, F., Kovac, J. & Greneche, J. M. Structural and magnetic properties of metastable  $\text{Fe}_{1-x}\text{Si}_x$  ( $0.15 < x < 0.34$ ) alloys prepared by a rapid-quenching technique. *J. Phys. Condens. Matter* **14**, 1985–2000 (2002).

**Acknowledgements** This research is supported by the Christian Doppler Laboratory for Thermoelectricity and the JST, CREST (grant numbers JPMJCR15Q6 and JPMJCR19Q4). Work at IMR, China, was supported by the National Science Fund for Distinguished Young Scholars (grant number 51725103), by the National Natural Science Foundation of China (grant number 51671193), by the Science Challenging Project (grant number TZ2016004), and by the Shanghai Nuclear Engineering Research & Design Institute (major research project 2018ZX06002004).

**Author contributions** B.H., I.K., M.P. and Y.S. contributed equally to the preparation of bulk materials and thin films, measurements and ab initio calculations. C.E.-S., B.H., H.M., G.E., M.S.-P., Y.K., N.K., Q.G., T.B., T.M. and E.B. collected and analysed the data. Y.S., S.U. and X.-Q.C. carried out the DFT calculations. All authors contributed to the interpretation of the data and to the writing of the manuscript

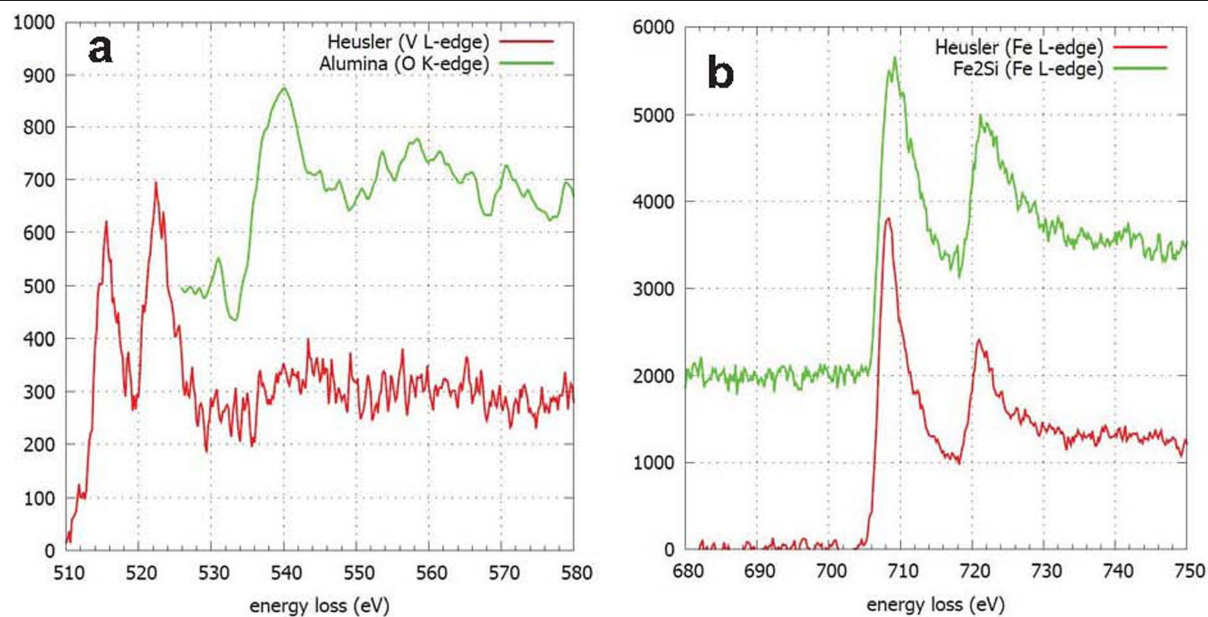
**Competing interests** The authors declare no competing interests.

#### Additional information

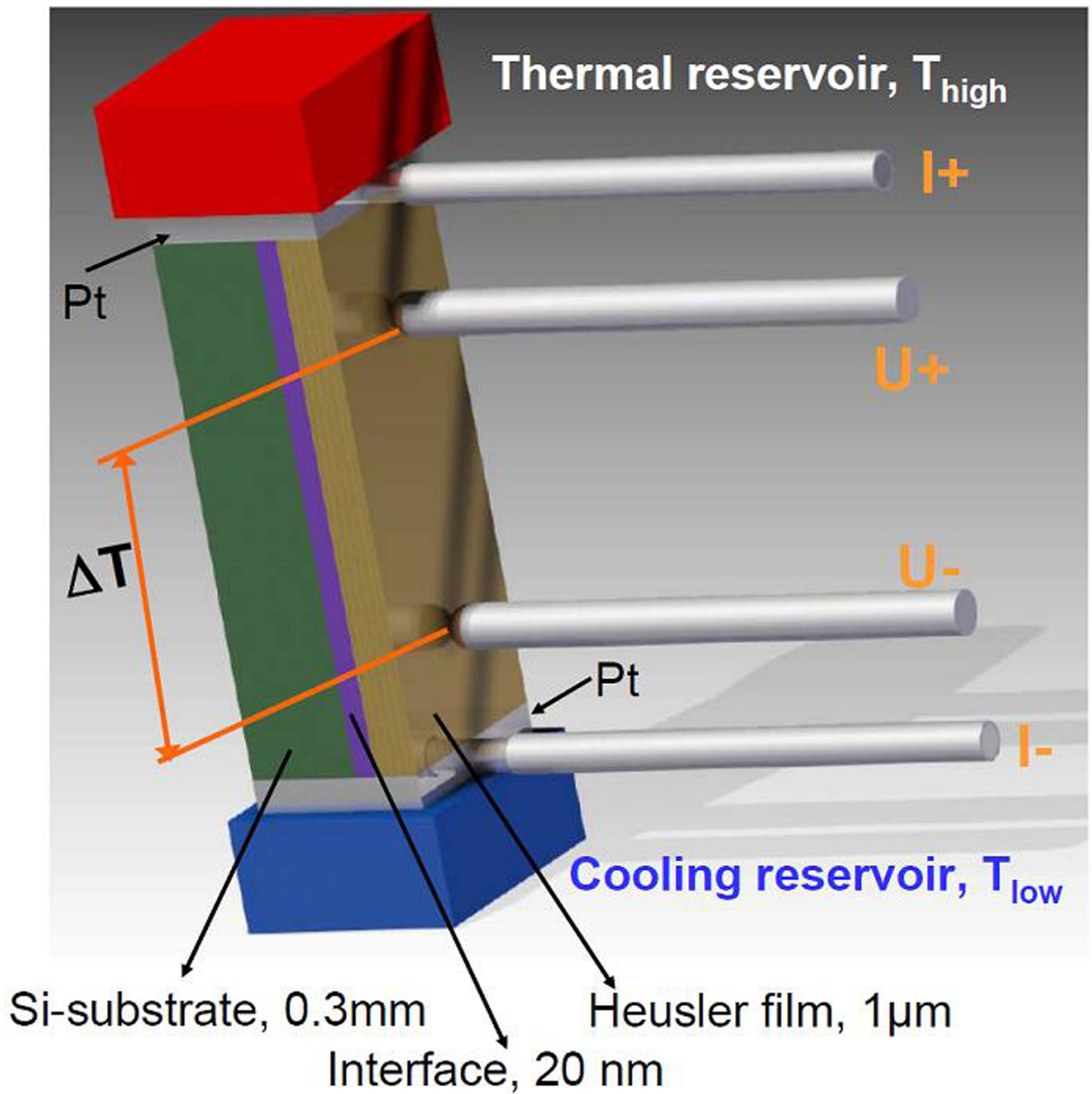
**Correspondence and requests for materials** should be addressed to E.B.

**Peer review information** Nature thanks Stephen R. Boona and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

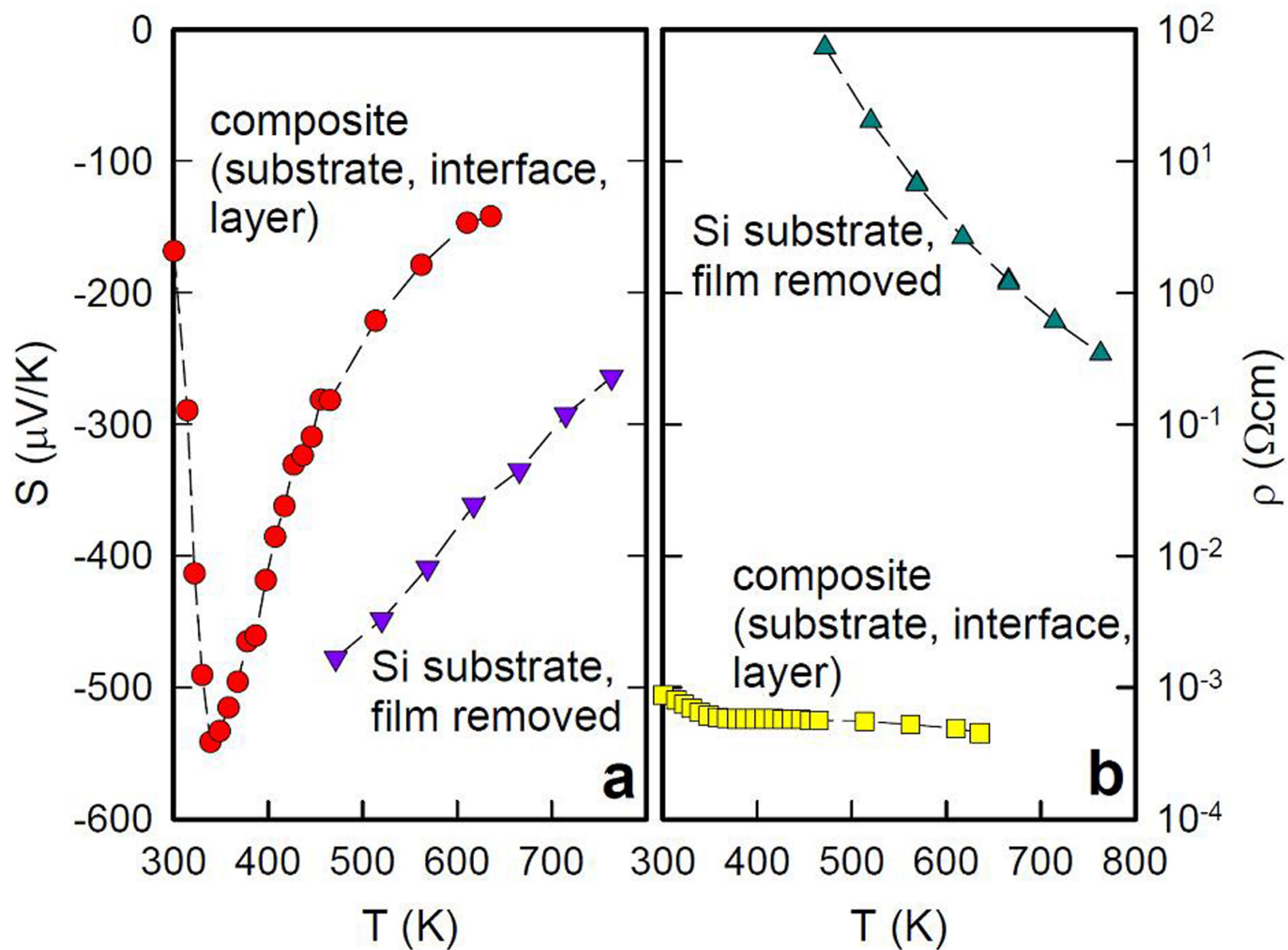
**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



**Extended Data Fig. 1 | Electron microscopy results at the interface between substrate and film. a,** ELNES of the oxygen K-edge of the thin alumina interlayer at 532-eV energy loss and the vanadium L-edge at 513-eV energy loss of the Heusler alloy. **b,** ELNES of the iron L-edge in Fe<sub>2</sub>Si and the Heusler alloy.

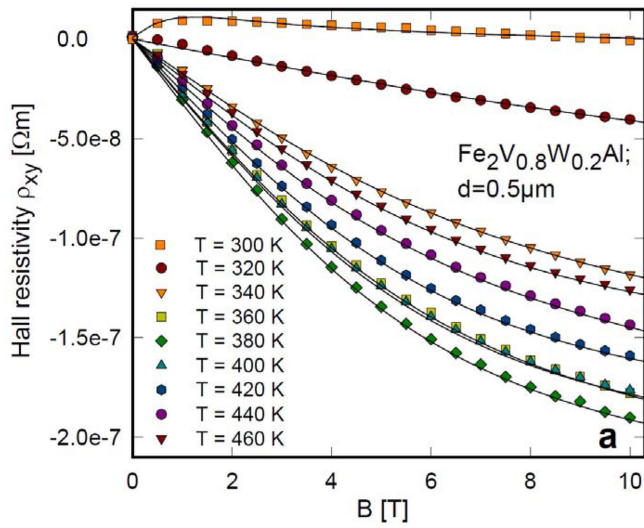


**Extended Data Fig. 2 | Schematic of the measurement set-up.** The thin film, the interface and the Si substrate are shown.  $\Delta T$  is the temperature difference between the electrodes, denoted as  $U^+$  and  $U^-$ .

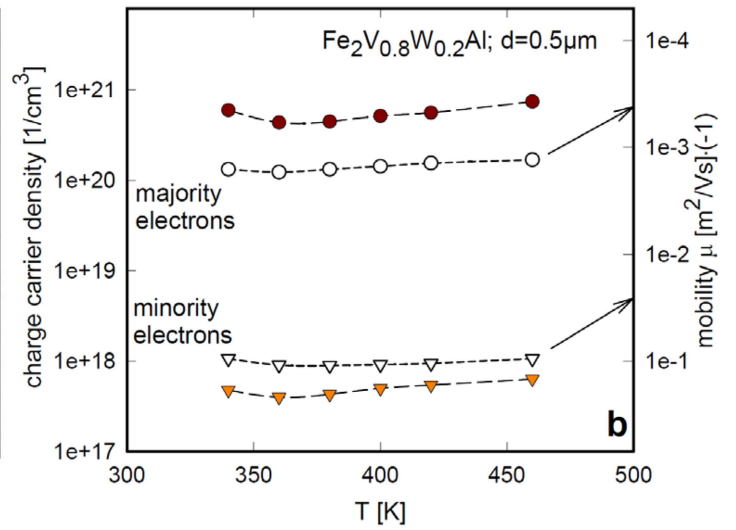


**Extended Data Fig. 3 | Electronic and thermoelectric transport of the composite (Heusler film, interface and Si substrate) and of the isolated Si substrate. a, Temperature-dependent Seebeck coefficient  $S$ . b, Temperature-dependent electrical resistivity  $\rho$ .**

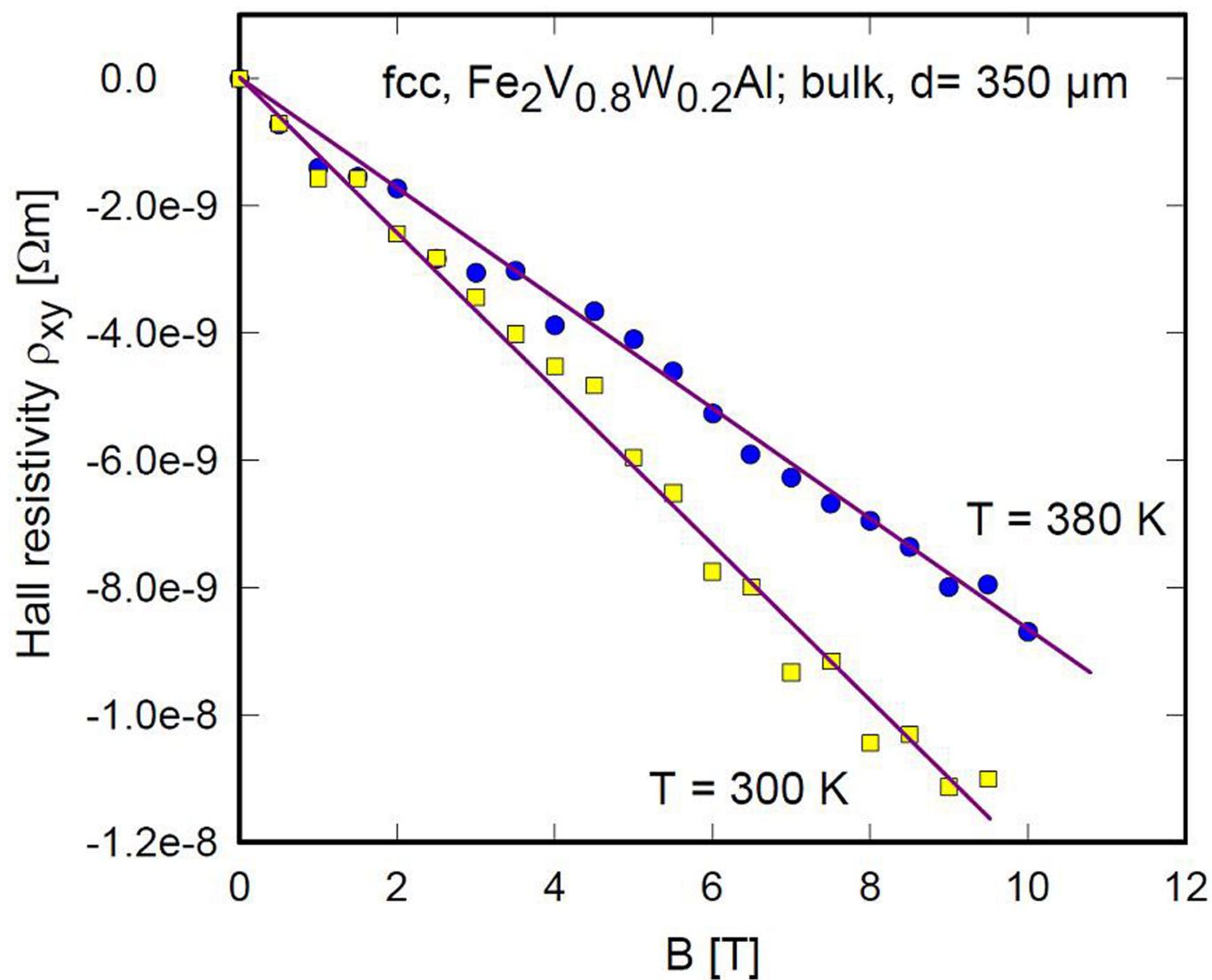




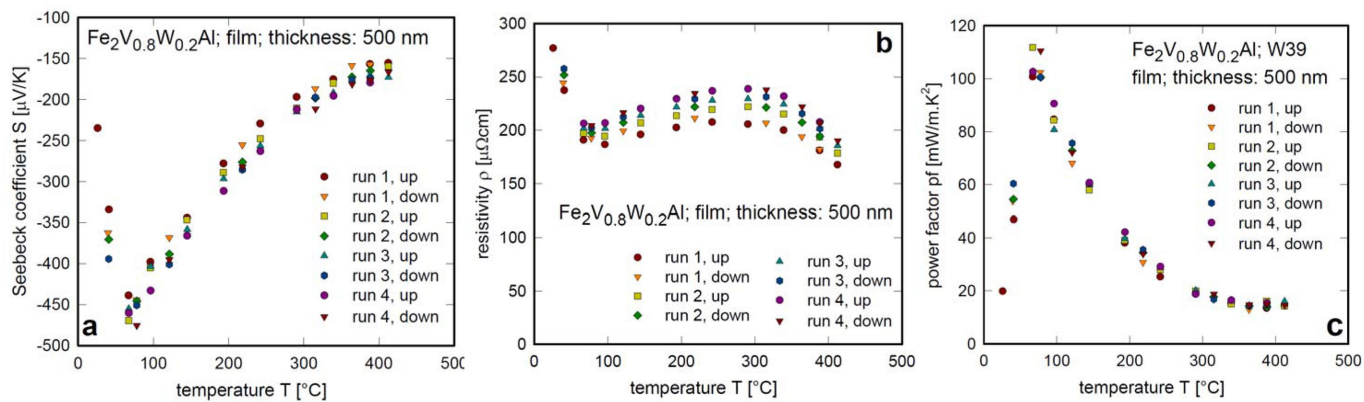
**Extended Data Fig. 4 | Field- and temperature-dependent Hall data of thin-film  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$ .** **a**, Field-dependent Hall resistivity  $\rho_{xy}$  at various temperatures. The solid lines are least-squares fits, as explained in Methods. **b**, Temperature-dependent charge carrier densities  $n$  and



mobilities  $\mu$ , derived from the above least-squares fits. Circle and triangle symbols indicate the data for the majority and minority electrons, respectively; the coloured and open symbols indicate the data for charge carrier density and mobility, respectively. The dashed lines are guides to the eye.

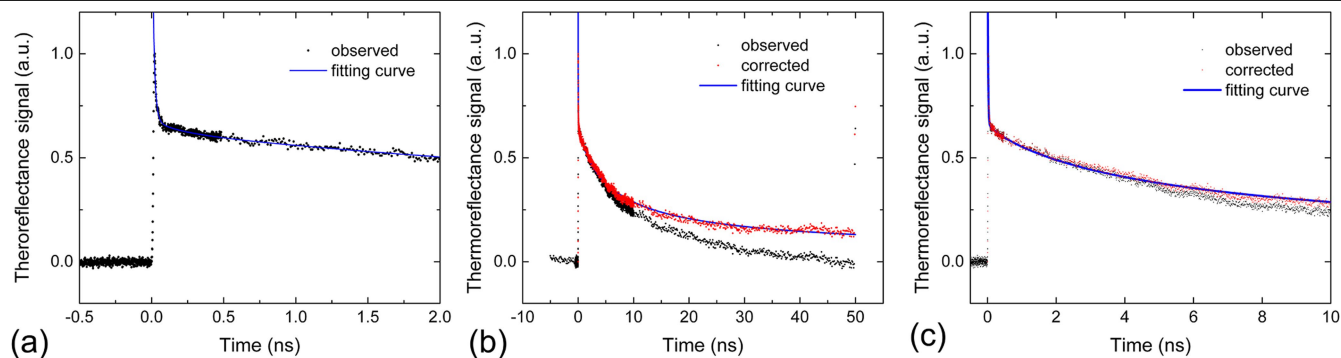


**Extended Data Fig. 5** | Field-dependent Hall resistivity  $\rho_{xy}$  of bulk  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  at  $T = 380 \text{ K}$  and  $T = 300 \text{ K}$ . The solid lines are least-squares fits, as explained in Methods.



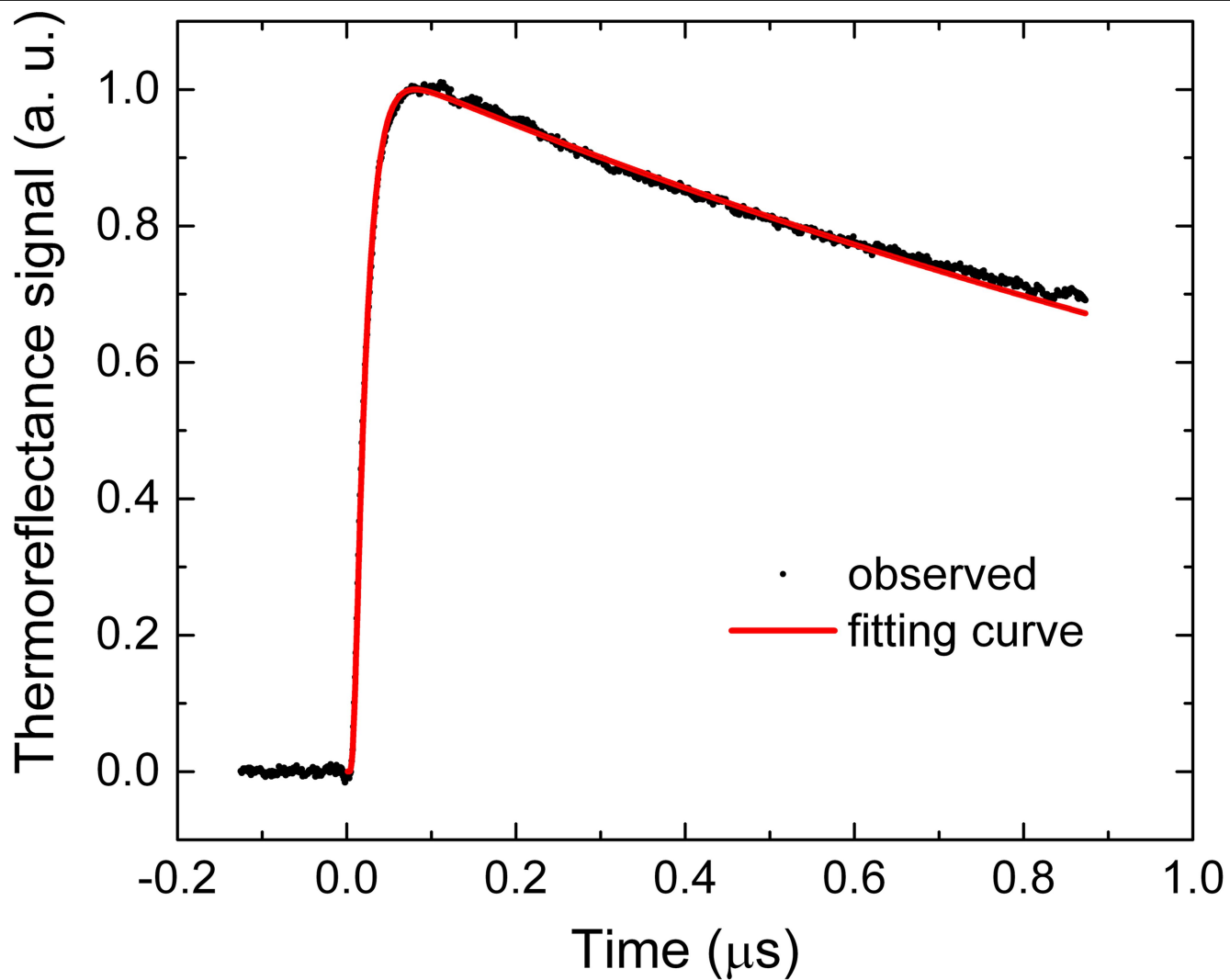
**Extended Data Fig. 6 | Temperature-dependent transport and thermoelectric properties of thin-film  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$ .** Data are taken over four runs, each for increasing and decreasing temperatures. **a**, Temperature-

dependent Seebeck coefficient  $S$ . **b**, Temperature-dependent electrical resistivity  $\rho$ . **c**, Temperature-dependent power factor PF.



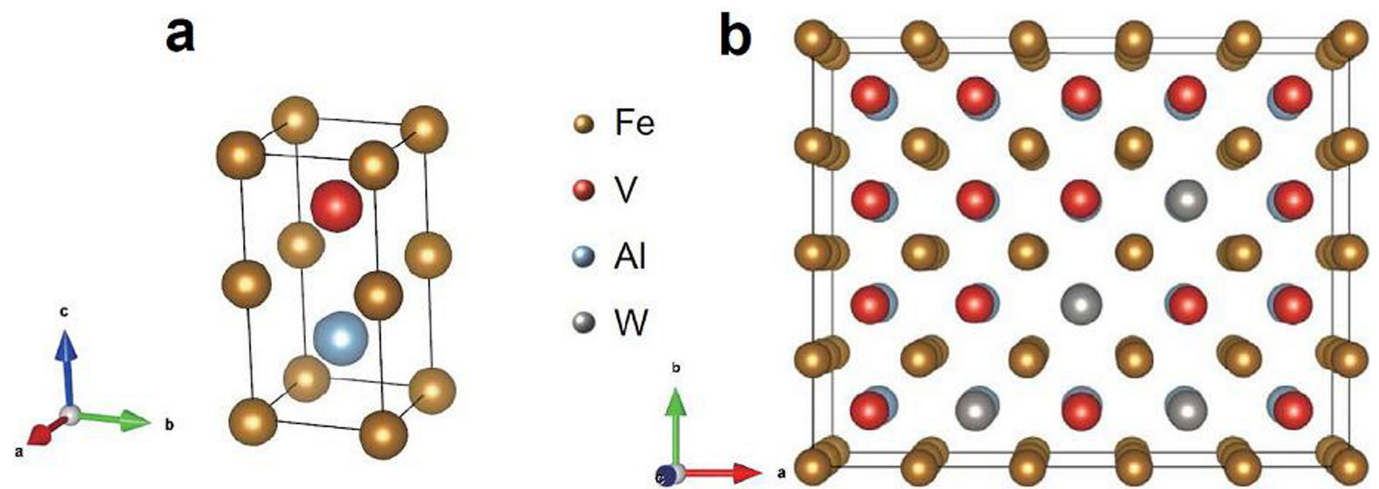
**Extended Data Fig. 7 | Thermoreflectance signal of thin film  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  with deposited 100-nm-thick Al surface layer.** The data are observed by the picosecond thermoreflectance method. **a**, The signal data are enlarged around the instant of pump laser irradiation at  $t = 0$  with the fitting curve calculated

using the mirror image method. **b**, The entire signal from pulse to pulse. **c**, The signal enlarged around the pump laser irradiation using the method reported in refs.<sup>54,55</sup>. The red dotted lines in **b** and **c** are corrected signals for single-pulse heating. The blue solid lines are least-squares fits as explained in Methods.

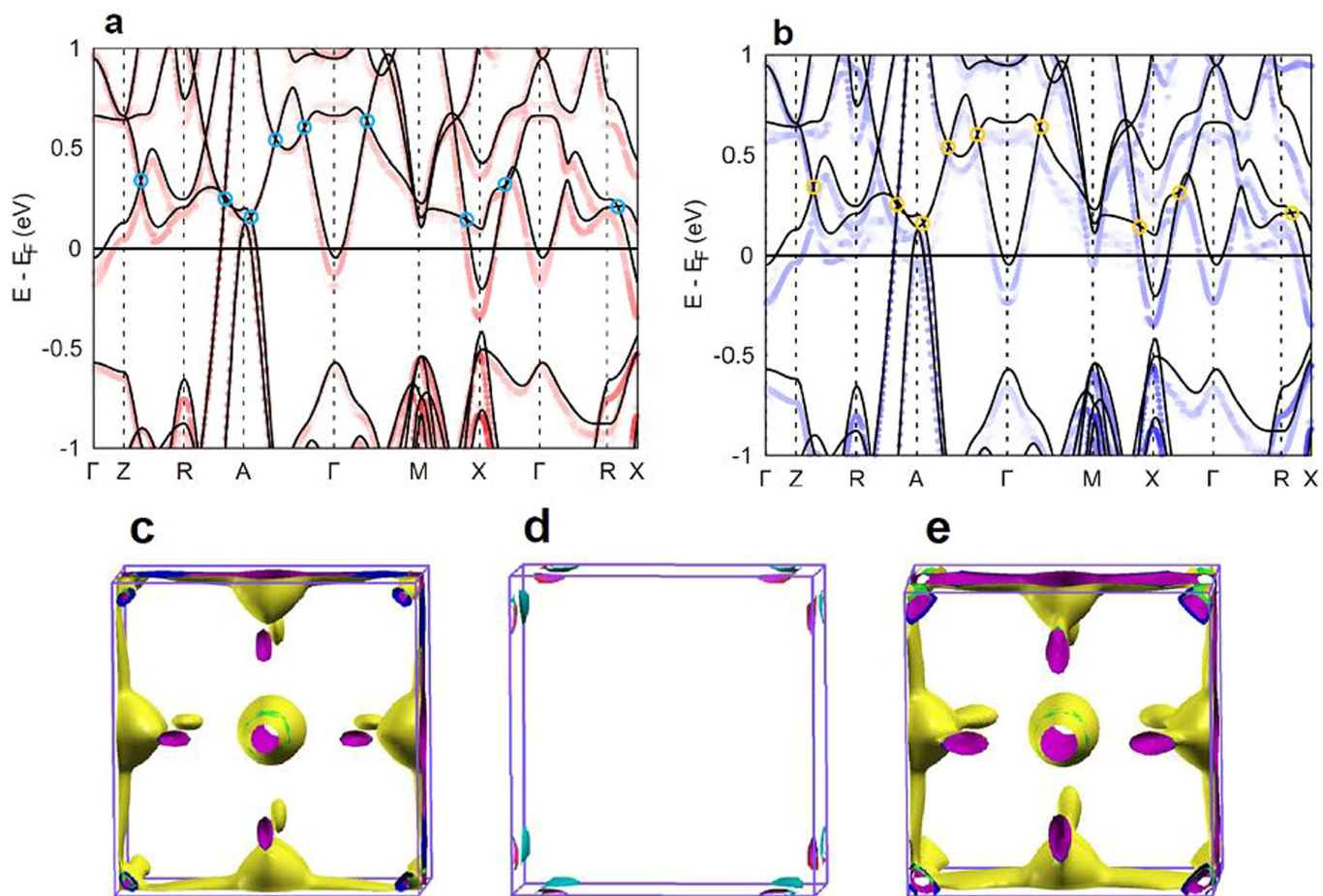


**Extended Data Fig. 8 | Thermoreflectance signal of the 680-nm-thick TiN reference film deposited on a quartz substrate.** The signal was obtained using the ultrafast laser flash method.





**Extended Data Fig. 9 | The lattice structure of full-Heusler compounds. a,** Parent material  $\text{Fe}_2\text{VAl}$ . **b,** Parent material  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$ . Descriptions and interpretations are summarized in Methods.



**Extended Data Fig. 10 | Electronic structure of bcc  $\text{Fe}_2\text{VAl}$  and  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$ .** **a, b**, Electron dispersion along the high-symmetry directions of the bcc structure around the Fermi level for spin-up (**a**) and spin-down bands (**b**). The dispersion of  $\text{Fe}_2\text{VAl}$  is illustrated by black solid lines; the coloured lines refer to

$\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$ . **c–e**, Fermi surfaces of bcc  $\text{Fe}_2\text{V}_{0.8}\text{W}_{0.2}\text{Al}$  at 0.11 eV above the Fermi level on Z–R (**c**); on R–A (**d**); and at 0.13 eV above the Fermi level on  $\Gamma$ –R (**e**). Descriptions and interpretations are summarized in Methods.

# Additive manufacturing of ultrafine-grained high-strength titanium alloys

<https://doi.org/10.1038/s41586-019-1783-1>

Received: 4 March 2019

Accepted: 8 October 2019

Published online: 4 December 2019

Duyao Zhang<sup>1,6</sup>, Dong Qiu<sup>1,6</sup>, Mark A. Gibson<sup>1,3</sup>, Yufeng Zheng<sup>2,5</sup>, Hamish L. Fraser<sup>2\*</sup>, David H. StJohn<sup>4</sup> & Mark A. Easton<sup>1\*</sup>

Additive manufacturing, often known as three-dimensional (3D) printing, is a process in which a part is built layer-by-layer and is a promising approach for creating components close to their final (net) shape. This process is challenging the dominance of conventional manufacturing processes for products with high complexity and low material waste<sup>1</sup>. Titanium alloys made by additive manufacturing have been used in applications in various industries. However, the intrinsic high cooling rates and high thermal gradient of the fusion-based metal additive manufacturing process often leads to a very fine microstructure and a tendency towards almost exclusively columnar grains, particularly in titanium-based alloys<sup>1</sup>. (Columnar grains in additively manufactured titanium components can result in anisotropic mechanical properties and are therefore undesirable<sup>2</sup>.) Attempts to optimize the processing parameters of additive manufacturing have shown that it is difficult to alter the conditions to promote equiaxed growth of titanium grains<sup>3</sup>. In contrast with other common engineering alloys such as aluminium, there is no commercial grain refiner for titanium that is able to effectively refine the microstructure. To address this challenge, here we report on the development of titanium–copper alloys that have a high constitutional supercooling capacity as a result of partitioning of the alloying element during solidification, which can override the negative effect of a high thermal gradient in the laser-melted region during additive manufacturing. Without any special process control or additional treatment, our as-printed titanium–copper alloy specimens have a fully equiaxed fine-grained microstructure. They also display promising mechanical properties, such as high yield strength and uniform elongation, compared to conventional alloys under similar processing conditions, owing to the formation of an ultrafine eutectoid microstructure that appears as a result of exploiting the high cooling rates and multiple thermal cycles of the manufacturing process. We anticipate that this approach will be applicable to other eutectoid-forming alloy systems, and that it will have applications in the aerospace and biomedical industries.

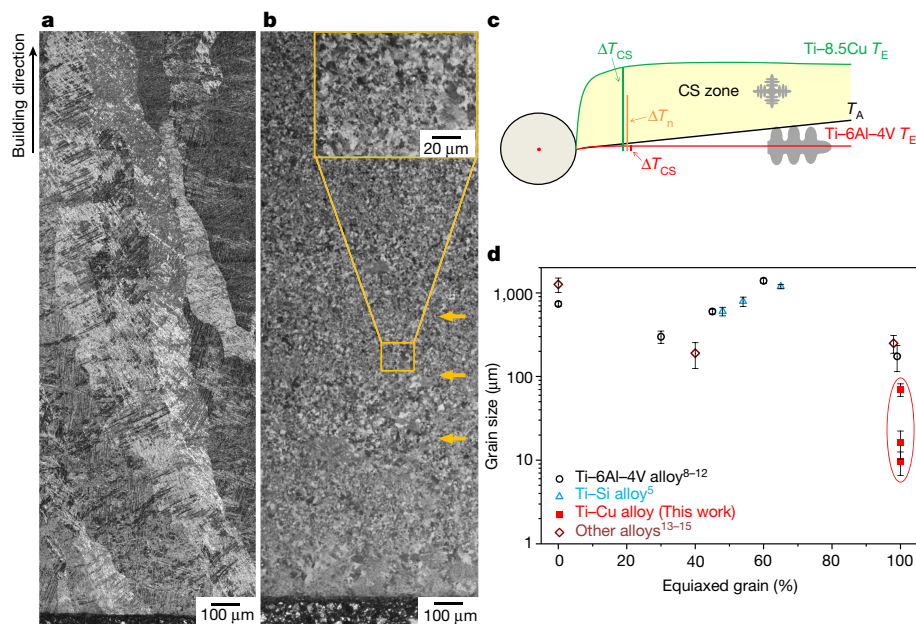
According to Interdependence Theory<sup>4</sup>, the key factors controlling grain size include: (1)  $\Delta T_n$ , the critical undercooling for nucleation; (2)  $\Delta T_{CS}$ , the amount of constitutional supercooling in front of the growing solid that provides the nucleation undercooling; and (3)  $x_{sd}$ , the average spacing between the potent nucleation particles. A small  $\Delta T_n$ , large  $\Delta T_{CS}$  and small  $x_{sd}$  favours grain refinement. The rate of development of a constitutional supercooling zone is controlled by the growth restriction factor  $Q$ . Larger values of  $Q$  promote more nucleation. However, in additively manufactured metals, the dimensions of the laser-melted region, coupled with a high thermal gradient, considerably suppress the extent of the constitutional supercooling zone making it challenging to achieve a fine grain size

in additively manufactured titanium alloys. Multiple research groups have explored the possibilities of adding solute elements such as beryllium, silicon or boron to stop epitaxial growth<sup>5</sup>. However, these solute elements only decrease the width of columnar grains of the additively manufactured titanium or only achieve a partial columnar-to-equiaxed transition. It hence remains an open question whether fully equiaxed grain structures in additively manufactured titanium alloys are practically achievable through conventional grain-refining paradigms.

It should be noted that in previous grain-refining studies, the normalized  $Q$  value— $m(k-1)$ , where  $m$  is the slope of the liquidus line and  $k$  is the solute partition coefficient—has frequently been used

<sup>1</sup>Centre for Additive Manufacturing, School of Engineering, RMIT University, Melbourne, Victoria, Australia. <sup>2</sup>Center for the Accelerated Maturation of Materials, Department of Materials Science and Engineering, The Ohio State University, Columbus, OH, USA. <sup>3</sup>The Commonwealth Scientific and Industrial Research Organisation (CSIRO) Manufacturing, Clayton, Victoria, Australia.

<sup>4</sup>School of Mechanical and Mining Engineering, University of Queensland, St Lucia, Queensland, Australia. <sup>5</sup>Present address: Department of Chemical and Materials Engineering, University of Nevada, Reno, Reno, NV, USA. <sup>6</sup>These authors contributed equally: Duyao Zhang, Dong Qiu. \*e-mail: [fraser.3@osu.edu](mailto:fraser.3@osu.edu); [mark.easton@rmit.edu.au](mailto:mark.easton@rmit.edu.au)

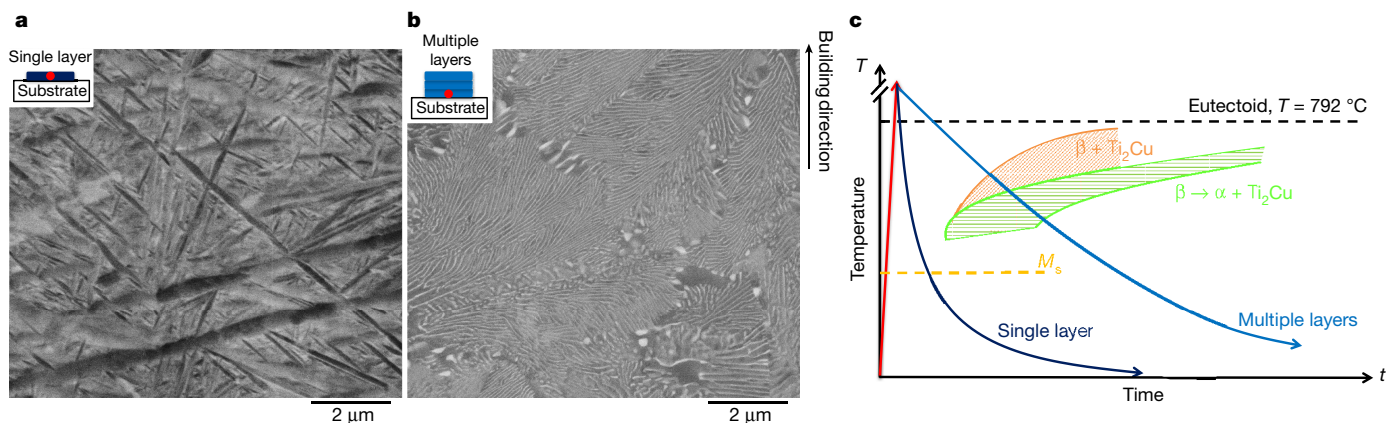


**Fig. 1 | Additive manufacturing of Ti-6Al-4V and Ti-8.5Cu alloys.** **a**, Optical micrograph of an as-printed Ti-6Al-4V alloy showing coarse columnar grains. **b**, By contrast, optical microstructures of an as-printed Ti-8.5Cu alloy show fine, fully equiaxed grains along the building direction under the same manufacturing conditions. The yellow arrows in **b** indicate successive layer boundaries approximately every 200  $\mu\text{m}$  and the average prior- $\beta$  grain size is 9.6  $\mu\text{m}$ , measured by the linear intercept technique. Inset, an enlarged portion of a local region with ultrafine grains. **c**, Schematic diagram of the grain growth mechanism of Ti-8.5Cu and Ti-6Al-4V alloys.  $T_A$  is the profile of the temperature of the melt and  $T_E$  is the profile of the equilibrium liquidus temperature. The values of  $\Delta T_{CS}(=T_E - T_A)$  and  $\Delta T_n$  are represented qualitatively

by the length bars. The red dot is the centre of the previous grain, which has grown to the size of the circle. The grey shapes represent the grain morphology for the two alloys. **d**, Summary of the area percentage of equiaxed grains versus grain size for the as-printed titanium alloys<sup>5,8–15</sup>. Ti-Si alloys (blue triangles) are, from top to bottom, Ti-0.04Si, Ti-0.19Si and Ti-0.75Si. The other titanium alloys (dark red diamonds) are, from left to right, Ti-6Al-2Zr-2Sn-3Mo-1Cr-2Nb, Ti-6.5Al-3.5Mo-1.5Zr-0.3Si and Ti-3Al-10V-2Fe. Most as-printed titanium alloys have either fully columnar or mixed columnar and equiaxed prior- $\beta$  grains and the grain sizes are in the range of 100  $\mu\text{m}$  to 1 mm. This work shows that fully equiaxed prior- $\beta$  grains can be achieved throughout the as-printed samples. Error bars represent one standard deviation.

to guide the choice of solute elements. However, the solubility of a given solute element in the  $\beta$ -phase titanium, which defines the practical maximum solute concentration,  $c_{0-\text{max}}$ , has been neglected. By simply exploring binary titanium alloy phase diagrams, we note copper to be a promising solute, with a  $c_{0-\text{max}}$  as high as 17 wt% and a reasonably high  $m(k-1)$  value of 6.5 K. This leads to an overall very high maximum  $Q$  value,  $Q_{\text{max}} = c_{0-\text{max}} m(k-1) = 110.5 \text{ K}$ , which far surpasses that of silicon or boron<sup>6</sup>.

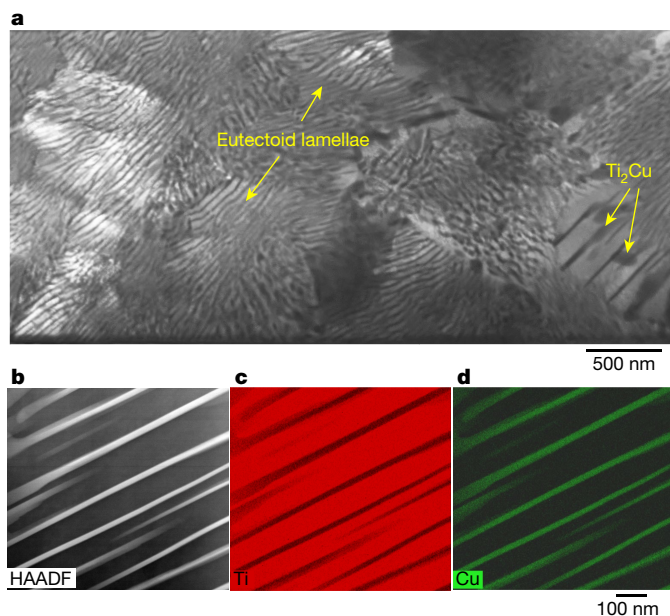
In addition to its potential for refining  $\beta$ -phase titanium grains, copper is also a typical eutectoid-forming element in titanium binary alloy systems where  $\beta \rightarrow \alpha + \text{Ti}_2\text{Cu}$  at 792  $^{\circ}\text{C}$ . Because copper diffuses rapidly in titanium, this eutectoid reaction cannot easily be prevented from occurring even after water quenching<sup>7</sup>. Such characteristics are beneficial to the high cooling rates during additive manufacturing and are likely to produce a very fine eutectoid microstructure, improving both the strength and ductility of as-printed specimens. Therefore,



**Fig. 2 | Scanning electron microscopy (SEM) characterization of Ti-8.5Cu alloy.** **a**, **b**, Backscattered electron (BSE) images of the as-printed Ti-8.5Cu alloy from Fig. 1b showing the microstructure evolution at the first layer (indicated by the red spots) during the additive manufacturing process, with constant processing parameters. The martensite phase forms when only a single layer was deposited (**a**); fine eutectoid lamellae surrounded by hyper-eutectoid

$\text{Ti}_2\text{Cu}$  particles form when multiple layers were deposited (**b**). **c**, A schematic continuous cooling transformation diagram illustrates different solid–solid phase transformation pathways for laser deposition of the first layer and the successive layers. Heat accumulates during the deposition of successive layers, thus the cooling rate is reduced and the  $\beta \rightarrow \alpha + \text{Ti}_2\text{Cu}$  reaction is complete before the martensite transformation temperature ( $M_s$ ) is reached.





**Fig. 3 | Transmission electron microscopy characterization of as-printed Ti-8.5Cu alloy.** **a**, Bright-field image showing the ultrafine eutectoid lamellar structure and a small portion of hyper-eutectoid  $\text{Ti}_2\text{Cu}$  particles close to the prior- $\beta$  grain boundaries. **b–d**, X-ray energy dispersive spectroscopy (XEDS) mapping on a section of the eutectoid lamellar structure: high-angle annular dark-field scanning transmission electron microscopy image (**b**), titanium elemental map (**c**) and copper elemental map (**d**). XEDS point analyses show that the copper contents in the lamellar structure are 2.8 wt% in  $\alpha$ -phase titanium and 39.1 wt% in  $\text{Ti}_2\text{Cu}$ . Under equilibrium conditions, the maximum solubility of copper in  $\alpha$ -phase titanium is 2.0 wt% and in  $\text{Ti}_2\text{Cu}$  it is 39.9 wt%<sup>21</sup>.

in the present study, we aim to develop additively manufactured titanium–copper alloys (Extended Data Fig. 1) to form fully equiaxed  $\beta$ -phase titanium grains and an ultrafine eutectoid microstructure in a one-step process.

The optical micrographs of the as-printed (see Methods) Ti-8.5Cu specimen (herein, we use weight per cent unless otherwise specified) show fully equiaxed prior- $\beta$  grains (primary Ti grains that form during solidification, as shown in Fig. 1b) without any noticeable cracks and with a small volume fraction of enclosed porosity (see Extended Data Fig. 2). The as-printed specimen also has excellent chemical homogeneity along the building direction (see Extended Data Fig. 3). The prior- $\beta$  grains have a bimodal distribution with an average grain size of 9.6  $\mu\text{m}$ . In comparison, the microstructure of as-printed Ti-6Al-4V alloy is dominated by coarse columnar grains (Fig. 1a) under the same laser processing conditions. It can be seen that the addition of copper has not only fully converted the columnar grains to equiaxed grains but also refined the prior- $\beta$  grains by two orders of magnitude. The commonly observed epitaxial growth is also completely eliminated, indicated by the size of the equiaxed grains, which is much smaller than the layer thickness of about 200  $\mu\text{m}$  (yellow arrows in Fig. 1b). It is also worth noting that compared with other additively manufactured titanium alloys reported thus far<sup>5,8–15</sup>, our current work has produced the smallest equiaxed prior- $\beta$  titanium alloy grains made by additive manufacturing, as shown in Fig. 1d. The grain-refining efficiency of the as-printed titanium–copper alloys stems from the high capacity of the copper solute to establish a sufficiently large constitutional supercooling zone in front of the solid–liquid interface, which is formed when the solute copper segregates around the first  $\beta$ -phase titanium dendritic grain (Fig. 1c); the  $Q$  value of the Ti-8.5Cu alloy is 62 K. By contrast, in Ti-6Al-4V, the Al and V solutes provide negligible constitutional supercooling (that is,  $Q = 8$  K), which is far less than the nucleation undercooling  $\Delta T_n$  during solidification. As a

result, wide columnar grains with an average width of 120  $\mu\text{m}$  grow in the Ti-6Al-4V alloy, but fine equiaxed grains of average dimension 9.6  $\mu\text{m}$  grow in the Ti-8.5Cu alloy. The constitutional supercooling,  $\Delta T_{cs}$ , is proportional to the  $Q$  value<sup>16</sup> through the dimensionless supersaturation parameter,  $\Omega$ :

$$\Delta T_{cs} = Q\Omega \quad (1)$$

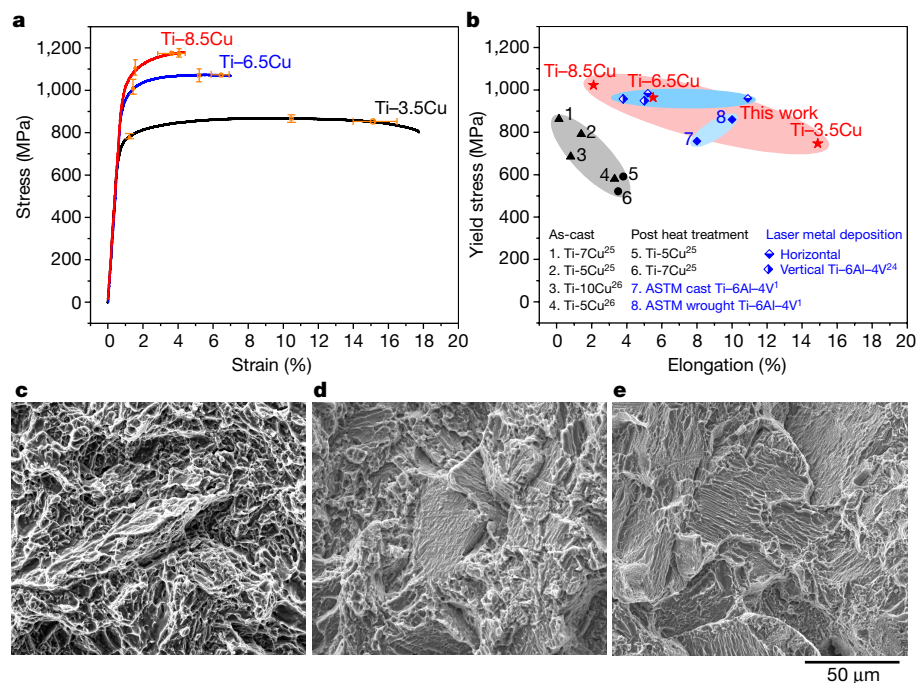
This means that the constitutional supercooling zone is eight times greater in magnitude during additive manufacturing of Ti-8.5Cu compared to Ti-6Al-4V, subjected to the same laser processing conditions. Sufficient constitutional supercooling can efficiently offset the negative impact of a high thermal gradient and ensures that waves of heterogeneous nucleation events can be triggered in the constitutional supercooling zone and a complete columnar-to-equiaxed transition can be achieved. By Interdependence Theory, the grain size is also dependent on  $Q$ . More copper solute delivers higher constitutional supercooling faster, and therefore the size of the equiaxed prior- $\beta$  grain is reduced with increasing copper content (see Extended Data Fig. 4).

It is worth mentioning that the Scheil–Gulliver solidification path and freezing range are often used to predict the likelihood of cracking during solidification<sup>17</sup>. A large freezing range usually leads to less liquid being available for interdendritic feeding during the last stage of solidification. In this study, Scheil curves show a large freezing range of more than 500 K (Extended Data Fig. 5, dashed line) based on the titanium–copper equilibrium phase diagram. However, no cracks in the as-printed titanium–copper specimens were observed. This can be at least partially explained by equation (1). As the required constitutional supercooling critical temperature for the columnar-to-equiaxed transition is usually between  $10^{-1}$  K to 10 K, the resultant supersaturation  $\Omega$  is much less than 1. This means that heterogeneous nucleation events occur very early during solidification. The formation of fine equiaxed dendrites can effectively decrease the hot-tearing susceptibility, as validated in previous studies of cast alloys<sup>18</sup>.

Upon completion of liquid-to- $\beta$ -phase solidification, the  $\beta$ -phase of titanium (a body-centred cubic structure) can decompose into different product phases in the subsequent solid–solid phase transformations subject to the cooling rate<sup>19</sup>. A high cooling rate can restrict the diffusion of atoms, which suppresses eutectoid coupled growth, resulting in martensite ( $\alpha'$ -phase titanium, hexagonal close-packed structure) formation<sup>20</sup>. Martensite in titanium alloys can lead to higher strength but lower ductility<sup>8</sup>. As expected, acicular plates of martensite (Fig. 2a) were observed as a result of the high cooling rate in the single track of the additively manufactured Ti-8.5Cu alloy; however, successive layer-by-layer fabrication leads to multiple thermal cycles above and below the eutectoid reaction temperature (792  $^{\circ}\text{C}$ ) in the previously deposited layer and thus the cooling rate of the  $\beta$ -phase decomposition decreases as the number of layers increases, owing to insufficient heat dissipation (see Fig. 2c). This characteristic thermal history can efficiently reverse the martensitic transformation and results in ultrafine eutectoid lamellae (Fig. 2b and Extended Data Fig. 6). Similar phenomena have been observed in other compositions as well (see Extended Data Fig. 7). Moreover, the average interlamellar spacing in the as-printed Ti-8.5Cu alloy is 46 nm  $\pm$  7 nm (Fig. 2b), which is much finer than conventionally manufactured water-quenched (about 150 nm) and furnace-cooled (about 1  $\mu\text{m}$ ) samples<sup>7</sup>. This is because the interlamellar spacing is controlled by the diffusion length of the copper atoms; the diffusion length is considerably restricted by fast cooling.

Titanium alloys, in general, have a very low thermal conductivity<sup>21</sup>,  $\leq 16 \text{ W m}^{-1} \text{ K}^{-1}$ , which may lead to interlamellar spacing coarsening from the surface to the core, owing to the variation in cooling rate during a conventional normalizing heat treatment for large, bulky





**Fig. 4 | Mechanical properties of as-printed Ti–Cu alloys.** **a**, Representative engineering stress–strain curves of the as-printed materials in this study; error bars represent one standard deviation. **b**, Yield strength (0.2% offset) versus tensile elongation to failure for Ti–Cu alloys manufactured by different methods<sup>24–26</sup>; the properties of these alloys are comparable with those of the

ASTM standard<sup>1</sup> for a Ti–6Al–4V alloy. **c**, Ductile fracture surface of Ti–3.5Cu showing small dimples. **d**, Fracture surface of Ti–6.5Cu showing a mixture of regions of small dimples with regions of cleavage facets. **e**, Brittle fracture surface of Ti–8.5Cu showing only cleavage facets.

titanium–copper components. By contrast, the laser metal deposition process enables relatively constant cooling rates across the alloy, leading to a more uniform microstructure regardless of the size of the specimen. Only a slight increase in interlamellar spacing from the bottom ( $41 \text{ nm} \pm 5 \text{ nm}$ ) to the top ( $54 \text{ nm} \pm 9 \text{ nm}$ ) of the specimen was observed (errors represent one standard deviation). This is a result of the probably decreased cooling rate along the building direction. It is also worth mentioning that the copper concentration in the eutectoid lamellae (Fig. 3b–d) deviates from the equilibrium composition. The  $\alpha$ -phase titanium contains 2.8 wt% copper and it is supersaturated, because the maximum solid solubility of copper in  $\alpha$ -phase titanium is 2.0 wt% at equilibrium. This indicates that a more substantial precipitation hardening effect could be achieved to further increase the tensile strength through optimized post heat treatment.

Tensile tests with subsized ASTM standard specimens were performed on the as-printed alloys and the associated 0.2% offset yield strength ( $\sigma_y$ ), ultimate tensile strength, and uniform elongation ( $\epsilon$ ) are summarized in Table 1. Comparing the Ti–6.5Cu and Ti–3.5Cu alloys, the eutectoid lamellae in Ti–6.5Cu increases the strength substantially but decreases the ductility (see Fig. 4a). Comparing the Ti–8.5Cu and Ti–6.5Cu alloys, Ti–8.5Cu has higher strength because of the higher volume fraction of eutectoid lamellae, but lower ductility owing to the hyper-eutectoid  $\text{Ti}_2\text{Cu}$  particles<sup>22</sup>

**Table 1 | Mechanical properties of as-printed Ti–Cu alloys**

Samples	$\sigma_y$ (MPa)	UTS (MPa)	$\epsilon$ (%)
Ti–3.5Cu	$747 \pm 7$	$867 \pm 8$	$14.9 \pm 1.9$
Ti–6.5Cu	$964 \pm 31$	$1073 \pm 27$	$5.5 \pm 0.4$
Ti–8.5Cu	$1023 \pm 29$	$1180 \pm 21$	$2.1 \pm 0.6$

(see Extended Data Table 1). The size of equiaxed prior- $\beta$  grains (Fig. 1b and Extended Data Fig. 4) and microstructural length-scales (Fig. 2b and Extended Data Fig. 7a, b) will probably also have an impact on the mechanical properties<sup>23</sup>. The fracture surfaces (Fig. 4c–e) show changes from dimples to a typical intragranular fracture morphology, which is consistent with the change in the ductility of the alloys. Compared with conventional casting and post-heat-treatment methods (Fig. 4b), the mechanical properties of the as-printed titanium–copper alloys with ultrafine equiaxed prior- $\beta$  grains and eutectoid lamellar structure display a superior combination of offset yield strength and ductility. The properties are also comparable to that of cast and wrought Ti–6Al–4V alloy<sup>1</sup>, as well as laser-metal-deposited Ti–6Al–4V alloy<sup>24</sup>. Furthermore, copper is a relatively low-cost alloying element and titanium–copper alloys can be additively manufactured with mixed elementary powders instead of with pre-alloyed powders. Titanium–copper alloys also have excellent antibacterial properties, good biocompatibility and corrosion resistance<sup>22,25–27</sup>. It is also anticipated that further improvements in a range of properties can be achieved through process manipulation using additive manufacturing.

We have demonstrated a pathway to additively manufacturing titanium–copper alloys with both fine equiaxed prior- $\beta$  grains and an ultrafine eutectoid lamellar structure. Our experimental results show that the solidification and subsequent eutectoid decomposition can be synergistically engineered to tailor mechanical properties to suit specific applications. This approach to grain refinement, using alloys with high  $Q$  values, has been demonstrated across many alloying systems<sup>28</sup> and solidification processes and has been demonstrated here as a design methodology for additively manufactured titanium alloys. The methodology is also likely to be applicable to other eutectoid systems such as for pearlitic steels, in which the mechanical properties of these conventional alloys could be enhanced by additive manufacturing for high-performance engineering applications.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1783-1>.

1. Zhang, D. et al. Metal alloys for fusion-based additive manufacturing. *Adv. Eng. Mater.* **20**, 1700952 (2018).
2. Carroll, B. E., Palmer, T. A. & Beese, A. M. Anisotropic tensile behavior of Ti–6Al–4V components fabricated with directed energy deposition additive manufacturing. *Acta Mater.* **87**, 309–320 (2015).
3. Herzog, D., Seyda, V., Wycisk, E. & Emmelmann, C. Additive manufacturing of metals. *Acta Mater.* **117**, 371–392 (2016).
4. StJohn, D. H., Qian, M., Easton, M. A. & Cao, P. The Interdependence Theory: the relationship between grain formation and nucleant selection. *Acta Mater.* **59**, 4907–4921 (2011).
5. StJohn, D. H. et al. The challenges associated with the formation of equiaxed grains during additive manufacturing of titanium alloys. *Key Eng. Mater.* **770**, 155–164 (2018).
6. Bermingham, M. J., McDonald, S. D., StJohn, D. H. & Dargusch, M. S. Beryllium as a grain refiner in titanium alloys. *J. Alloys Compd.* **481**, L20–L23 (2009).
7. Cardoso, F. F. et al. Hexagonal martensite decomposition and phase precipitation in Ti–Cu alloys. *Mater. Des.* **32**, 4608–4613 (2011).
8. Xu, W., Lui, E. W., Pateras, A., Qian, M. & Brandt, M. In situ tailoring microstructure in additively manufactured Ti–6Al–4V for superior mechanical performance. *Acta Mater.* **125**, 390–400 (2017).
9. Mitzner, S., Liu, S., Domack, M. S. & Hafley, R. A. Grain refinement of freeform fabricated Ti6Al4V alloy using beam/arc modulation. In *23rd Solid Freeform Fabrication Symp.* 536–555 (2012); <https://sffsymposium.engr.utexas.edu/Manuscripts/2012/2012-42-Mitzner.pdf>.
10. Wang, F., Williams, S. & Rush, M. Morphology investigation on direct current pulsed gas tungsten arc welded additive layer manufactured Ti6Al4V alloy. *Int. J. Adv. Manuf. Technol.* **57**, 597–603 (2011).
11. Mereddy, S. et al. Trace carbon addition to refine microstructure and enhance properties of additive-manufactured Ti–6Al–4V. *JOM* **70**, 1670–1676 (2018).
12. Wang, J. et al. Grain morphology evolution and texture characterization of wire and arc additive manufactured Ti–6Al–4V. *J. Alloys Compd.* **768**, 97–113 (2018).
13. Li, Z., Li, J., Zhu, Y., Tian, X. & Wang, H. Variant selection in laser melting deposited  $\alpha + \beta$  titanium alloy. *J. Alloys Compd.* **661**, 126–135 (2016).
14. Zhu, Y.-Y., Tang, H.-B., Li, Z., Xu, C. & He, B. Solidification behavior and grain morphology of laser additive manufacturing titanium alloys. *J. Alloys Compd.* **777**, 712–716 (2019).
15. Zhu, Y., Liu, D., Tian, X., Tang, H. & Wang, H. Characterization of microstructure and mechanical properties of laser melting deposited Ti–6.5Al–3.5Mo–1.5Zr–0.3Si titanium alloy. *Mater. Des.* **56**, 445–453 (2014).
16. Kurz, W. & Fisher, D. J. *Fundamentals of Solidification* 3rd edn (Trans Tech Publications, 1989).
17. Fulcher, B. A., Leigh, D. K. & Watt, T. J. Comparison of AlSi10Mg and Al 6061 processed through DMLS. In *Proc. Solid Freeform Fabrication (SFF) Symp.* **46**, 404–419 (2014).
18. Easton, M., Wang, H., Grandfield, J., StJohn, D. & Sweet, E. An analysis of the effect of grain refinement on the hot tearing of aluminium alloys. *Mater. Forum* **28**, 224–229 (2004).
19. Souza, S. A., Afonso, C. R. M., Ferrandini, P. L., Coelho, A. A. & Caram, R. Effect of cooling rate on Ti–Cu eutectoid alloy microstructure. *Mater. Sci. Eng. C* **29**, 1023–1028 (2009).
20. Williams, J. C., Taggart, R. & Polonis, D. H. The morphology and substructure of Ti–Cu martensite. *Metall. Trans.* **1**, 2265–2270 (1970).
21. Brandes, E. A. & Brook, G. B. *Smithells Metals Reference Book* 7th edn (Butterworth-Heinemann, 1992).
22. Zhang, E., Wang, X., Chen, M. & Hou, B. Effect of the existing form of Cu element on the mechanical properties, bio-corrosion and antibacterial properties of Ti–Cu alloys for biomedical application. *Mater. Sci. Eng. C* **69**, 1210–1221 (2016).
23. Ren, Y. M. et al. Microstructure and deformation behavior of Ti–6Al–4V alloy by high-power laser solid forming. *Acta Mater.* **132**, 82–95 (2017).
24. Kok, Y. et al. Anisotropy and heterogeneity of microstructure and mechanical properties in metal additive manufacturing: a critical review. *Mater. Des.* **139**, 565–586 (2018).
25. Hayama, A. O. F. et al. Effects of composition and heat treatment on the mechanical behavior of Ti–Cu alloys. *Mater. Des.* **55**, 1006–1013 (2014).
26. Kikuchi, M. et al. Mechanical properties and microstructures of cast Ti–Cu alloys. *Dent. Mater.* **19**, 174–181 (2003).
27. Liu, R. et al. Antibacterial effect of copper-bearing titanium alloy (Ti–Cu) against *Streptococcus mutans* and *Porphyromonas gingivalis*. *Sci. Rep.* **6**, 29985 (2016).
28. Easton, M. A., Qian, M., Prasad, A. & StJohn, D. H. Recent advances in grain refinement of light metals and alloys. *Curr. Opin. Solid State Mater. Sci.* **20**, 13–24 (2016).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

## Methods

### Laser metal deposition

Pure (99.9%) titanium and (99.5%) copper spherical powders (TLS Technik and Thermo Fisher, respectively) with diameters between 50  $\mu\text{m}$  and 100  $\mu\text{m}$  (see Extended Data Fig. 8) were blended in a Turbula shaker mixer for an hour to achieve the designed compositions. Laser metal deposition was performed on a Trumpf TruLaser cell 7020. Before manufacturing bulk samples, we used single-layer deposition to optimize the processing parameters on the basis of visual observations of the weld bead. The optimized laser metal deposition parameters for the studied alloys are: laser power, 800 W; scanning speed, 800  $\text{mm min}^{-1}$ ; laser spot size, 1.5 mm; powder flow rate, 2.0  $\text{rpm}$  (1.7  $\text{g min}^{-1}$ ); hatch distance, 1.05 mm; shielding gas (argon) flow, 16  $\text{l min}^{-1}$ . The processing parameters were kept the same for all Ti–xCu alloys. Three cubes of  $10 \times 10 \times 10 \text{ mm}^3$  were built on a commercially pure titanium plate with different compositions (3.5 wt%, 6.5 wt% and 8.5 wt% copper). The laser scanning route for laser metal deposition was a raster pattern with an increment of  $90^\circ$  between each layer and the delay time between two subsequent layers was 20 s. For comparison, a Ti–6Al–4V specimen was additively manufactured using the same parameters.

For the tensile samples, three cuboids ( $120 \times 25 \times 25 \text{ mm}^3$ ) were horizontally built and then machined into five tensile samples. The loading direction of tensile samples is perpendicular to the laser metal deposition building direction.

### Chemical compositions

The chemical composition of the as-printed samples was determined by inductively coupled plasma-atomic emission spectroscopy, as summarized in Extended Data Table 1. A small amount of copper evaporation is expected, as its boiling point is 2,560  $^\circ\text{C}$ , much lower than that of titanium (3,285  $^\circ\text{C}$ )<sup>21</sup>.

### X-ray micro computed tomography

The as-printed samples were scanned using X-ray micro computed tomography (GE Phoenix V tome) with a nominal resolution of 5  $\mu\text{m}$ . Defect analysis including 3D image reconstruction, relative density, dimension and percentage of the defects was performed with Volume Graphics software. The as-printed specimens are all fully dense (>99.4%) without any lack of fusion defects. The porosity found in the as-printed specimens may come from the existing porosity in the powders (see Extended Data Fig. 8) as well as the manufacturing process.

### X-ray diffraction

Phase identification was determined by X-ray diffraction (XRD) using a Bruker AXS D4 Endeavour diffractometer over a  $2\theta$  range between  $15^\circ$  and  $90^\circ$  at a scanning rate of 0.06  $^\circ \text{s}^{-1}$ .

### Microscopy

As-printed cube samples were cut along the central section parallel to the build direction. All the samples as well as raw powders were prepared by mounting, grinding and polishing. The samples for optical microscopy were etched by Kroll's reagent to reveal the grain boundaries. Light optical microscopy with a polarized lens was used for examination of the microstructure. SEM in BSE mode was carried out using a FEI Verios 460L. Fracture topography was analysed using SEM in secondary electron mode.

The average grain size was measured from five optical micrographs of each alloy using the linear intercept technique. Volume fraction of lamellae phase was calculated from three BSE microstructure images ( $5,000\times$  magnification) by using the colour threshold.

For transmission electron microscopy (TEM) sample preparation, SEM with a focused ion beam (FEI Scios) was used to prepare site-specific TEM foils. Then, scanning transmission electron microscopy and XEDS mapping was performed in an image-corrected Titan3 G2 60-300 (S)TEM equipped with FEI's ChemiSTEM technology.

### Solidification simulation

Equilibrium and Scheil–Gulliver solidification models were simulated using Pandat software with PanTitanium database (version 2018). The  $Q$  values for Ti–6Al–4V and Ti–8.5Cu were determined from Scheil cooling curves<sup>29</sup>.

### Tensile testing

As-printed samples were machined into rectangular tension test specimens with gauge length of 25 mm and thickness of 4 mm (sub-size specimen of ASTM standard E8/E8 M-08). The tensile test loading direction is perpendicular to the laser metal deposition building direction. Quasi-static uniaxial testing was carried out at room temperature with an initial strain rate of  $1.0 \times 10^{-3} \text{ s}^{-1}$  on a universal testing facility (MTS810, 100 kN) equipped with a non-contact laser extensometer. Five tensile specimens were tested for each composition (see Extended Data Fig. 9). The results were then compared with ASTM standards for standard-size specimens.

### Data availability

The datasets generated or analysed during the current study are available from the corresponding author on reasonable request.

29. Schmid-Fetzer, R. & Kozlov, A. Thermodynamic aspects of grain growth restriction in multicomponent alloy solidification. *Acta Mater.* **59**, 6133–6144 (2011).
30. Okamoto, H. *Phase Diagrams For Binary Alloys* 2nd edn (ASM International, 2010).

**Acknowledgements** We acknowledge the Australian Research Council (ARC) for financial support (grant number DP160100560). D.Q. would like to thank the RMIT Vice-Chancellor's Senior Research Fellowship Fund for support. We thank M. Brandt and A. Jones for their support during laser metal deposition manufacturing, K. Yang for her support in etching additively manufactured titanium samples and E. Lui for his support in tensile testing. We acknowledge the facilities, and the scientific and technical assistance, of the RMIT Microscopy and Microanalysis Facility (RMMF). We also acknowledge the Center for Electron Microscopy and Analysis (CEMAS) at the Ohio State University for providing access to research facilities.

**Author contributions** M.A.E., M.A.G., D.H.StJ. and H.L.F. conceived the idea. D.Q. and D.Z. designed the experiments. D.Z. and D.Q. helped with processing parameter development and sample manufacturing, and performed the microstructure characterization. Y.Z. and D.Z. conducted the transmission electron microscopy and analysed the results. D.Z. performed mechanical testing, simulations and X-ray computed tomography. M.A.E. supervised the project. D.Z. and D.Q. drafted the manuscript. All authors discussed the results and edited the manuscript at all stages.

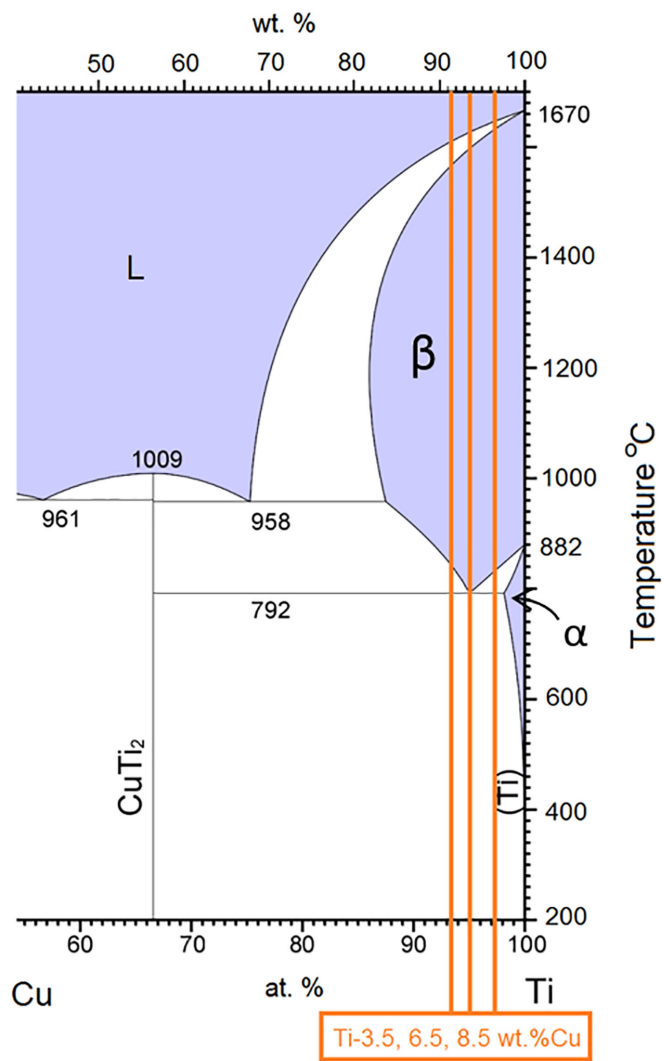
**Competing interests** The authors declare no competing interests.

### Additional information

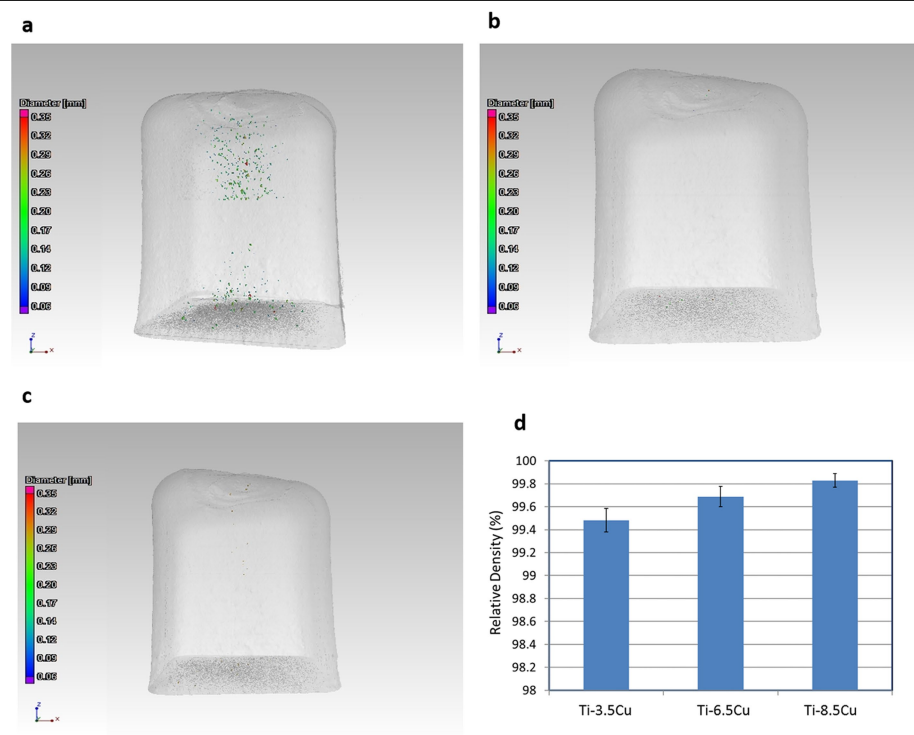
**Correspondence and requests for materials** should be addressed to H.L.F. or M.A.E.

**Peer review information** Nature thanks Amy Clarke, David Dye and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

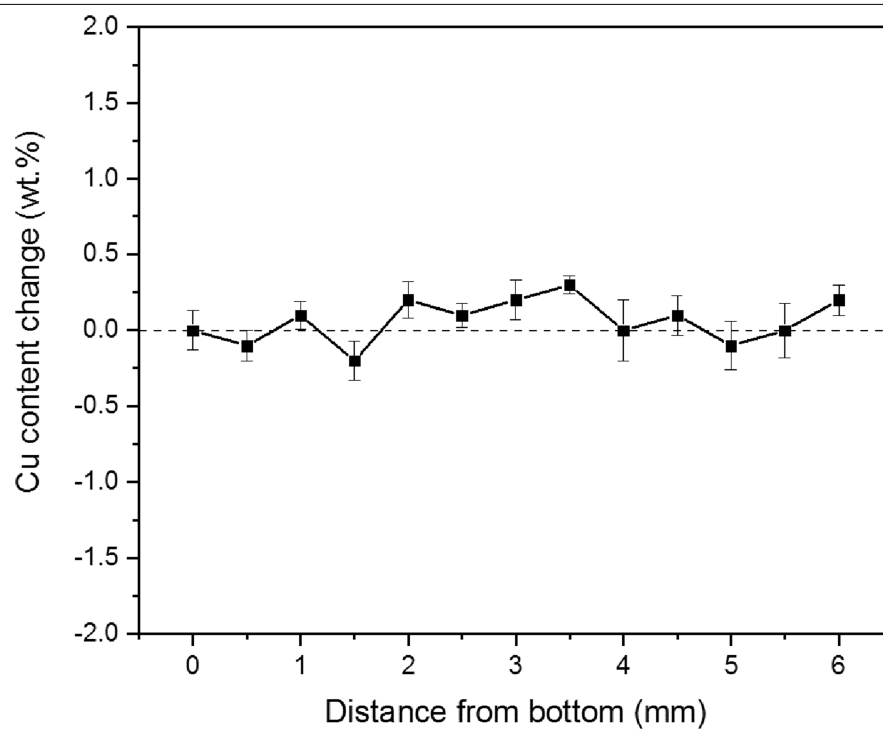


**Extended Data Fig. 1 | Ti-Cu phase diagram.** Portion of the Ti-Cu phase diagram indicating the compositions selected for laser metal deposition. We selected 3.5, 6.5 and 8.5 wt% copper to explore the behaviour of hypo-eutectoid, eutectoid and hyper-eutectoid compositions under additive manufacturing. This figure is adapted from ref.<sup>30</sup>, with the permission of ASM International.

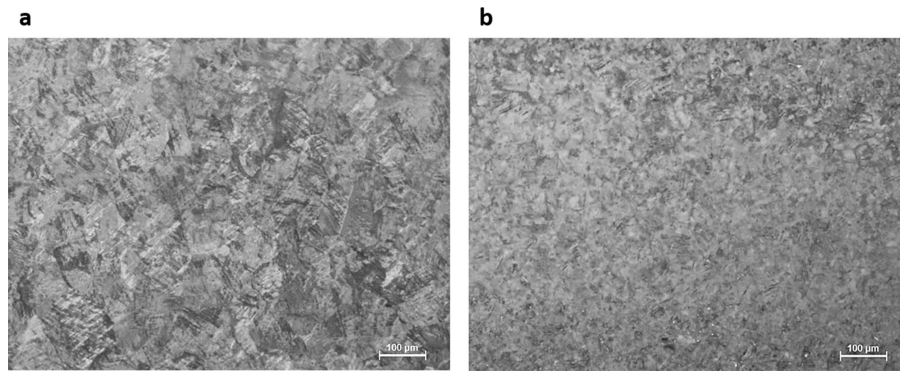


**Extended Data Fig. 2 | 3D visualization of the porosity of the manufactured specimens in the xyz coordinate system. a, Ti-3.5Cu. b, Ti-6.5Cu. c, Ti-8.5Cu. d, Calculated relative density of the as-printed specimens. Error bars represent one standard deviation.**

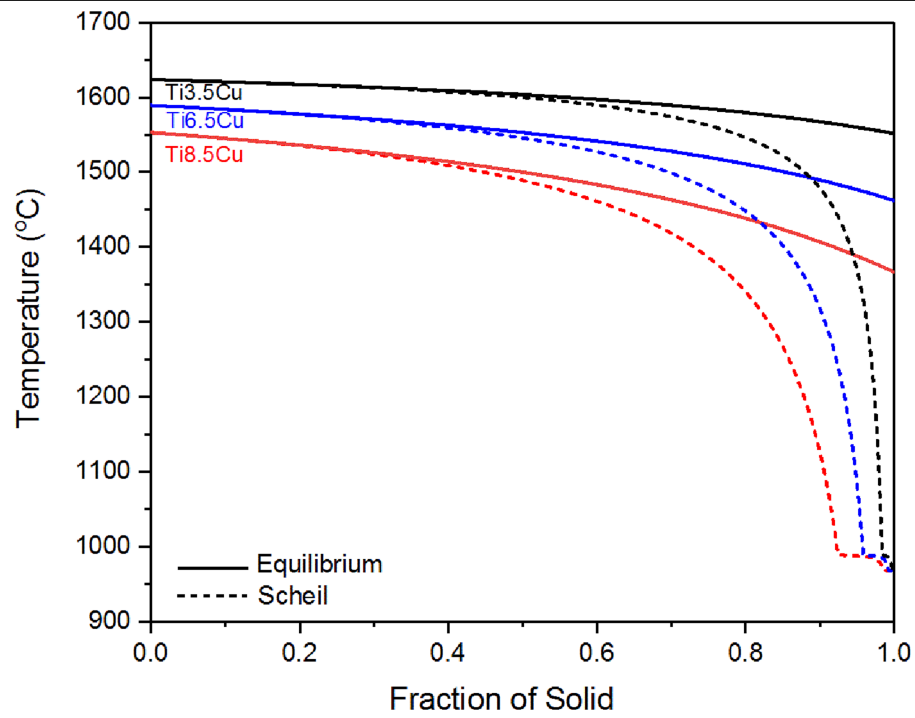




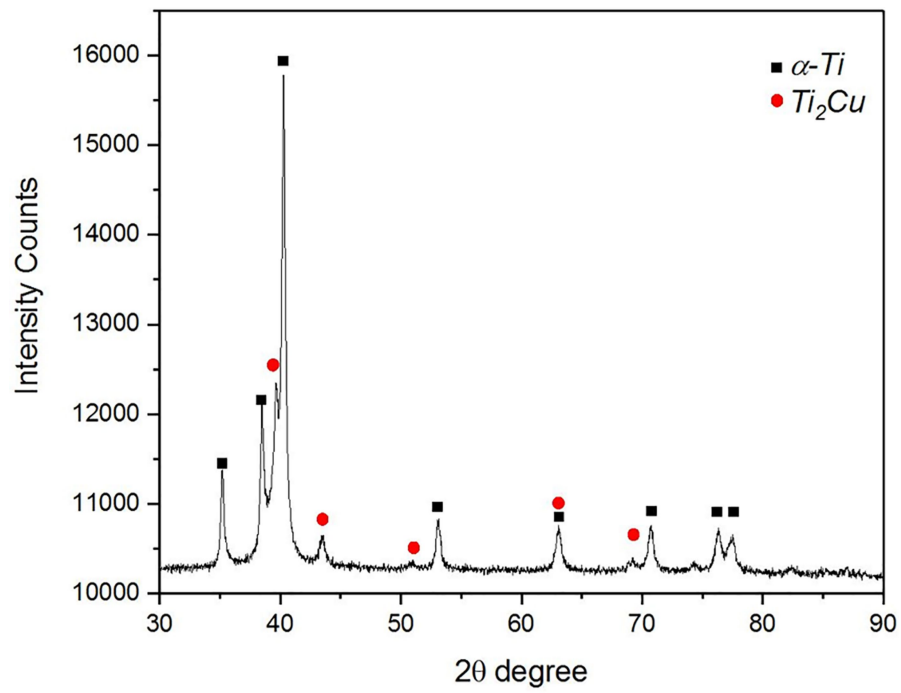
**Extended Data Fig. 3 | XEDS results of the copper content along the building direction for Ti-8.5Cu alloy.** The base point is 0 mm and the chemical composition is homogeneous. Error bars represent one standard deviation.



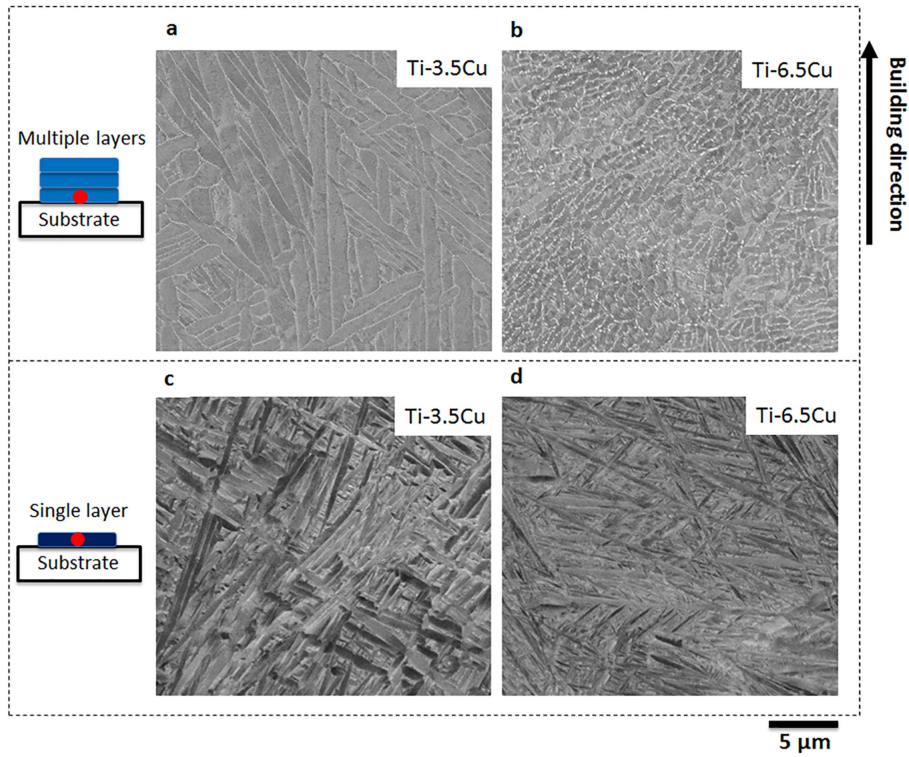
**Extended Data Fig. 4 | Polarized optical microstructures. a, b,** The equiaxed grains of as-printed Ti-3.5Cu (**a**) and Ti-6.5Cu (**b**). The average grain size is 69.8  $\mu\text{m}$  for Ti-3.5Cu and 16.3  $\mu\text{m}$  for Ti-6.5Cu.



**Extended Data Fig. 5 | Solidification curves.** The data are shown for different copper compositions under equilibrium and Scheil conditions. The Scheil curves show a substantially enlarged temperature interval between liquidus and solidus temperatures compared with the equilibrium condition.



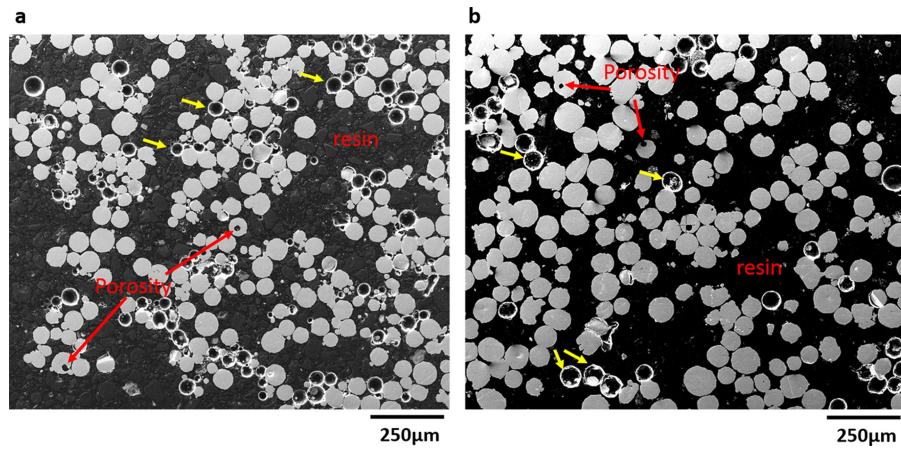
**Extended Data Fig. 6 | XRD spectra.** Experimental XRD spectra collected from the as-printed Ti-8.5Cu alloy indicates that only two phases are present in the specimen:  $\alpha$ -phase titanium and  $Ti_2Cu$ .



**Extended Data Fig. 7 | BSE images. a–d**, BSE images of as-printed specimens showing the fine  $\alpha$  phases when multiple layers were deposited, for Ti-3.5Cu (a) and Ti-6.5Cu (b); and the martensite phase when only a single layer was

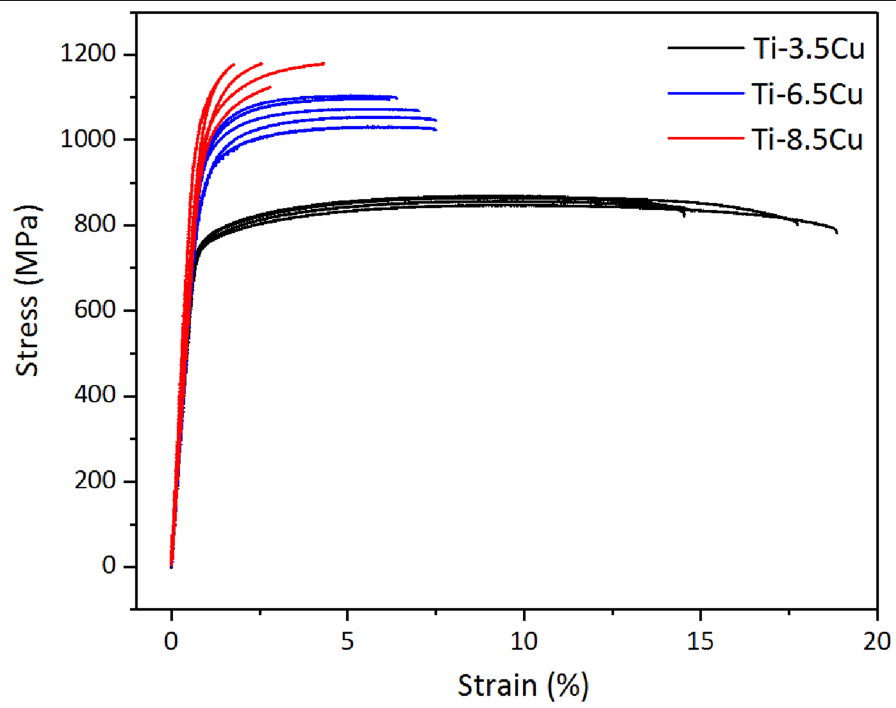
deposited for Ti-3.5Cu (c) and Ti-6.5Cu (d). Images were taken at the first layer of build specimens, indicated by the red spots.





**Extended Data Fig. 8 | SEM images of the cross-section of raw powders.**  
**a, b,** SEM images of the titanium powder (**a**) and copper powder (**b**) cross-sections. The powders are spherical in shape with a diameter between 50 μm

and 100 μm, and porosity can be observed within some powder particles. The yellow arrows indicate examples where powder particles fell out of the resin during the polishing process.



**Extended Data Fig. 9 | Engineering stress-strain curves.** The data for the additively manufactured materials tested in this study indicate good repeatability.

Extended Data Table 1 | Measured chemical compositions (wt%) and volume fraction of eutectoid lamellae in the as-printed alloys

Alloy (nominal composition)	Cu	N	O	Ti	Eutectoid lamellae (%)
Ti-3.5Cu	3.20	0.01	0.22	Bal.	0
Ti-6.5Cu	6.33	0.02	0.23	Bal.	53 ± 7
Ti-8.5Cu	8.36	0.02	0.21	Bal.	92 ± 4

Errors represent one standard deviation.

# Upper-plate rigidity determines depth-varying rupture behaviour of megathrust earthquakes

<https://doi.org/10.1038/s41586-019-1784-0>

Received: 21 June 2019

Accepted: 20 September 2019

Published online: 27 November 2019

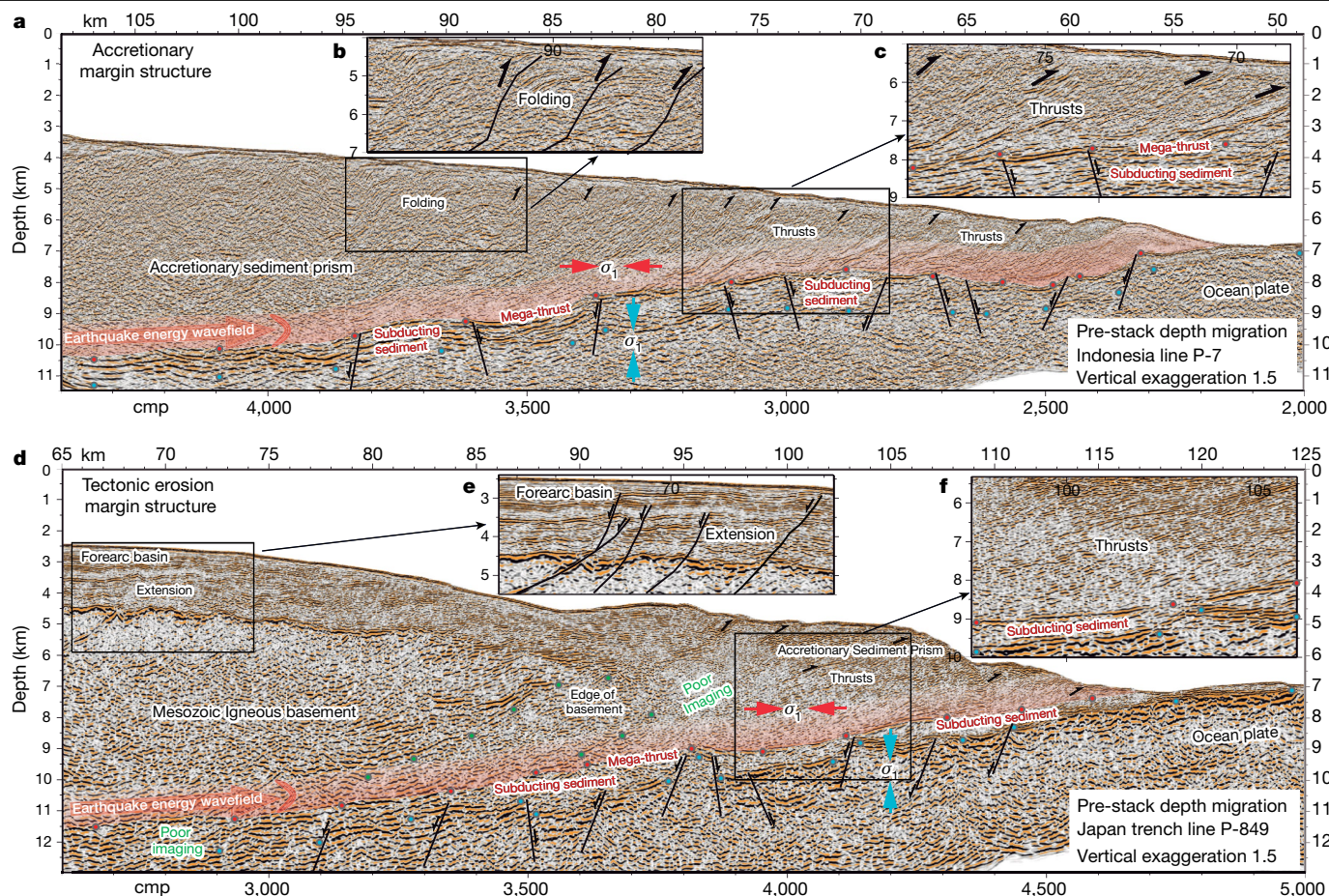
Valentí Sallarès<sup>1\*</sup> & César R. Ranero<sup>1,2</sup>

Seismological data provide evidence of a depth-dependent rupture behaviour of earthquakes occurring at the megathrust fault of subduction zones, also known as megathrust earthquakes<sup>1</sup>. Relative to deeper events of similar magnitude, shallow earthquake ruptures have larger slip and longer duration, radiate energy that is depleted in high frequencies and have a larger discrepancy between their surface-wave and moment magnitudes<sup>1–3</sup>. These source properties make them prone to generating devastating tsunamis without clear warning signs. The depth-dependent rupture behaviour is usually attributed to variations in fault mechanics<sup>4–7</sup>. Conceptual models, however, have so far failed to identify the fundamental physical causes of the contrasting observations and do not provide a quantitative framework with which to predict and link them. Here we demonstrate that the observed differences do not require changes in fault mechanics. We use compressional-wave velocity models from worldwide subduction zones to show that their common underlying cause is a systematic depth variation of the rigidity at the lower part of the upper plate – the rock body overriding the megathrust fault, which deforms by dynamic stress transfer during co-seismic slip. Combining realistic elastic properties with accurate estimates of earthquake focal depth enables us to predict the amount of co-seismic slip (the fault motion at the instant of the earthquake), provides unambiguous estimations of magnitude and offers the potential for early tsunami warnings.

Subduction megathrust earthquakes result from episodic, unstable sliding within the seismogenic zone<sup>8</sup>, a fault segment that is thought to extend from about 40–50 km to about 5–10 km depth. Great earthquakes initiating within the seismogenic zone can propagate updip from this limit, as evidenced for the 2011 Tohoku-Oki event<sup>9</sup> (moment magnitude,  $M_w$ , of 9.1) and 2010 Maule event ( $M_w$ , 8.8)<sup>10</sup>, while events forming a particular class known as ‘tsunami earthquakes’ appear to rupture only the shallowest, allegedly non-seismogenic part of the megathrust<sup>11</sup> (Extended Data Fig. 1). The seemingly anomalous characteristics of shallow ruptures suggest a depth dependency of the rupture process<sup>1–3</sup>, commonly attributed to changes in fault properties<sup>4–7</sup>. However, current conceptual models trying to explain the differences are qualitative and case-dependent; they treat the different rupture characteristics individually, as if they were caused by unrelated factors, and do not pinpoint the primary physical causes. Slow rupture propagation<sup>12,13</sup> and large slip<sup>14,15</sup>, for instance, are commonly attributed to the presence of weak subducting sediment<sup>16</sup>, whereas pore-pressure-related weakening<sup>4,5</sup> and a depth-dependent distribution of initial stresses<sup>6</sup> have also been proposed to explain large slip and high-frequency depletion. None of these models has been used to explain the remarkable discrepancy between  $M_w$  and surface-wave magnitude,  $M_s$ , for shallow earthquakes.

We propose a conceptual change to this unsolved question. Our hypothesis is that changes in fault mechanics are not necessarily required to explain the observed depth-dependent trends of the rupture characteristics. Instead, we postulate that the trend mainly reflects depth variations of the elastic properties of the overriding plate at a larger scale. This hypothesis stands on the fact that downgoing oceanic slabs and overriding plates exhibit contrasting patterns of permanent deformation<sup>17,18</sup> (Fig. 1). Overriding plates display widespread contractional structures indicating a dominant sub-horizontal principal compressional stress, whereas oceanic plates are dominated by extensional faulting, implying a near 90° rotation of the orientation of the principal stresses across the megathrust. Sedimentary strata of underthrusting plates have sub-horizontal attitude, typically lack contractional deformation and are cut by normal faults, supporting the idea that the principal compressional stresses are sub-vertical immediately below the megathrust fault. Thus, the distribution of tectonic structures and the inferred orientation of principal stresses support the idea that the elastic energy released during megathrust earthquakes has accumulated in overriding plates (Fig. 1). Correspondingly, co-seismic deformation should affect overriding plates, with negligible effect on the underthrusting plates. Hence, the recorded tectonic history indicates that the elastic properties of the overriding plate need to be considered

<sup>1</sup>Barcelona Center for Subsurface Imaging, Institute of Marine Sciences, CSIC, Barcelona, Spain. <sup>2</sup>ICREA, Barcelona, Spain. \*e-mail: vsallares@icm.csic.es



**Fig. 1 | Tectonic structure of the shallow region of two types of subduction zones where tsunamis are generated. a–c,** Depth-migrated multi-channel seismic image of the Java Trench (a). The overriding plate is made of accreted sediment thrust sheets, with different structure where the prism is more than about 5 km thick (inset b), and in the front, where it is less than about 5 km thick (inset c). Thrust at the front gradually rotated as material accumulated, thickening the prism landward (c), but when thrusts are too steep to continue slipping, out-of-sequence thrusts and folding further thicken and compact the prism (b). **d–f,** Pre-stack depth migration of the Japan Trench dominated by tectonic erosion<sup>17</sup>. The igneous basement flexes accumulating elastic energy and is cut by normal faults in its upper section (inset e; no vertical exaggeration). The frontal ~25 km is a sediment prism less than about 5 km thick, with thrust faulting (inset f; no vertical exaggeration). Both margins (a, d)

show contraction structures in the overriding plate indicating sub-horizontal main compressional stress. However, under the mega-thrust fault, the downgoing plate displays a fundamentally different structure: The top of the oceanic igneous crust is traced from the incoming plate into the underthrusting slab, overlaid by a layer of little-deformed sediment strata. The oceanic plate is characterized by horst and graben associated to bend-faulting, indicating a sub-vertical main stress. We interpret that the properties of the rock body deforming during rupture propagation (in red in the images) should change considerably within about the frontal 50 km of the margin. Stresses will be transferred through relatively consolidated material at about 10 km depth to progressively more fractured material at about 5 km depth, and a highly disaggregated upper plate, in the thinnest frontal 15 km or so of the overriding plate. cmp, common mid-point. P-7 and P-849 indicate the line profiles shown.

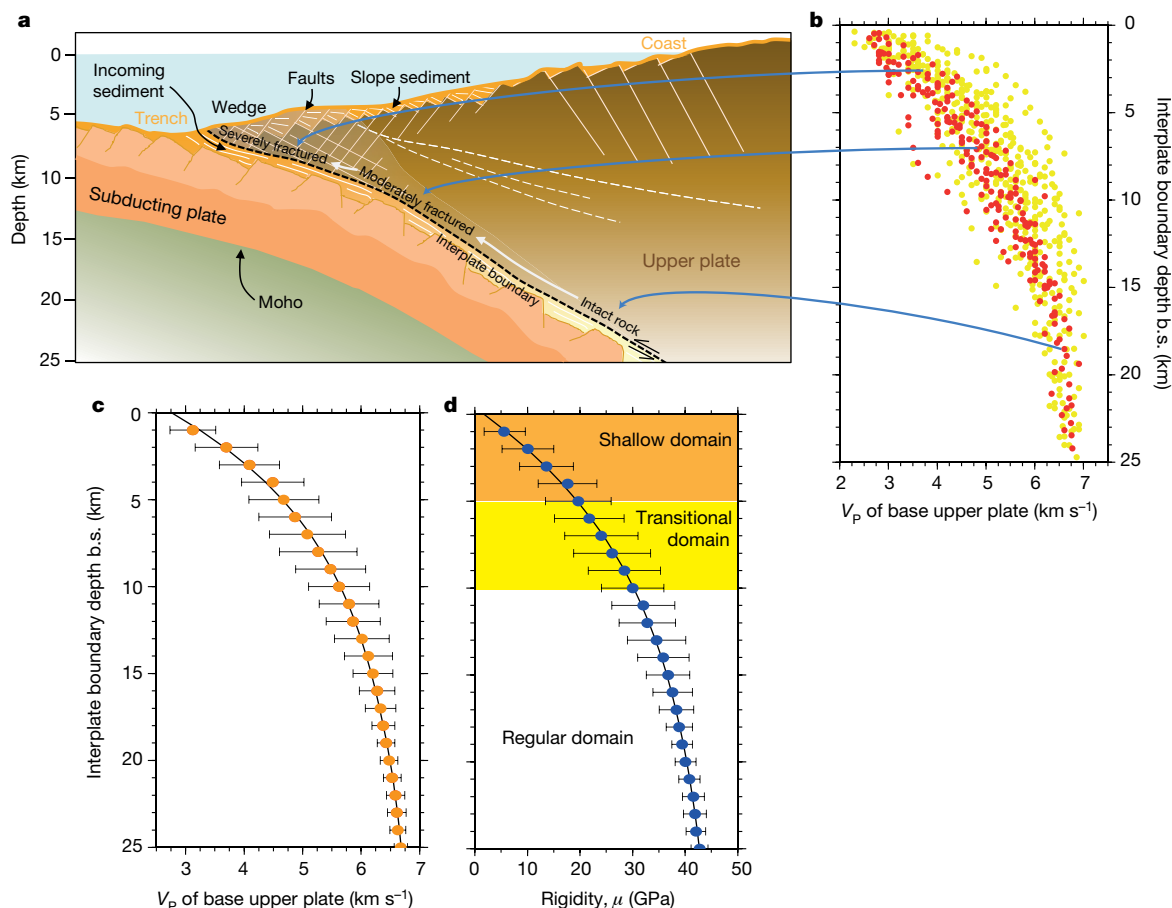
to understand the earthquake phenomena, given the constraints that they impose on dynamic stress transfer during co-seismic slip.

Our hypothesis implies that differences in rupture behaviour should be predictable and quantifiable if the depth distribution of elastic properties is known, and this information could be used to improve tsunami hazard assessment. To test it, we used 48 compressional-wave velocity ( $V_p$ ) models obtained with travel-time modelling of wide-angle reflection and refraction seismic profiles across circum-Pacific and Indian Ocean subduction zones (Extended Data Fig. 1 and Extended Data Table 1). We averaged  $V_p$  at the lower part of the overriding plate as a function of interplate boundary depth below seafloor, from the surface to approximately 25 km depth (Fig. 2 and Methods). The travel-time-based  $V_p$  models allow the rock volume encompassing the propagating rupture front to be resolved (Extended Data Fig. 3 and Methods). The global  $V_p(z)$  trends of accretionary and erosional margins, where  $z$  is interplate boundary depth below the seafloor, display slight differences at depths shallower than 5 km and gradually converge below this depth (Fig. 2b and Extended Data Fig. 4a).  $V_p(z)$  variations probably reflect

differences in rock nature between the two margin types in the shallow part, and a progressive rock compaction and a decrease in fracturing at deeper levels, as suggested by seismic images (Fig. 1). On average,  $V_p$  increases by a factor of 2.0–2.5, from about 3.0 km s<sup>-1</sup> at 1 km depth to about 6.5 km s<sup>-1</sup> at 25 km depth (Fig. 2c), with gradient decreasing downwards (Extended Data Fig. 5a).

We then use  $V_p$  to derive rigidity ( $\mu = \rho V_s^2$ , where  $\rho$  is density and  $V_s$  is shear-wave velocity), which affects important aspects of earthquake rupture. In the absence of direct  $V_s(z)$  measurements, we apply well-established empirical relationships for  $\rho(V_p)$  and  $V_s(V_p)$  from experimental data on multiple rock types<sup>19</sup> (see Methods). The resulting distributions of  $\rho(z)$  and  $V_s(z)$  are shown in Extended Data Fig. 3, and that for  $\mu(z)$  is shown in Fig. 2d. The trend shows a 4–5-fold increase in  $\mu$  between the surface and about 25 km depth, from <10 GPa at 1–2 km depth to 40–45 GPa at 20–25 km depth, with gradient decreasing downwards (Extended Data Fig. 5b). Based on the observed rates of variation, we define three domains along the megathrust: shallow (0–5 km), transitional (5–10 km) and regular (10–25 km) (Fig. 2d).





**Fig. 2 | Convergent margin structure and P-wave velocity at the lower part of overriding plates.** **a**, Conceptual model with main geological features of convergent margins, based on geophysical data of Central and South America. Decrease in  $V_p$  and intensified faulting towards the trench is interpreted to reflect increasing porosity and fracturing degree (see also Fig. 1). **b**, Coloured circles show digitized  $V_p$  values of the lower part of the upper plate just above the interplate boundary, as a function of interplate boundary depth below seafloor (depth b.s.) ( $z$ ). Red circles correspond to accretionary margins, yellow circles to erosional margins. Location of profiles in the compilation is

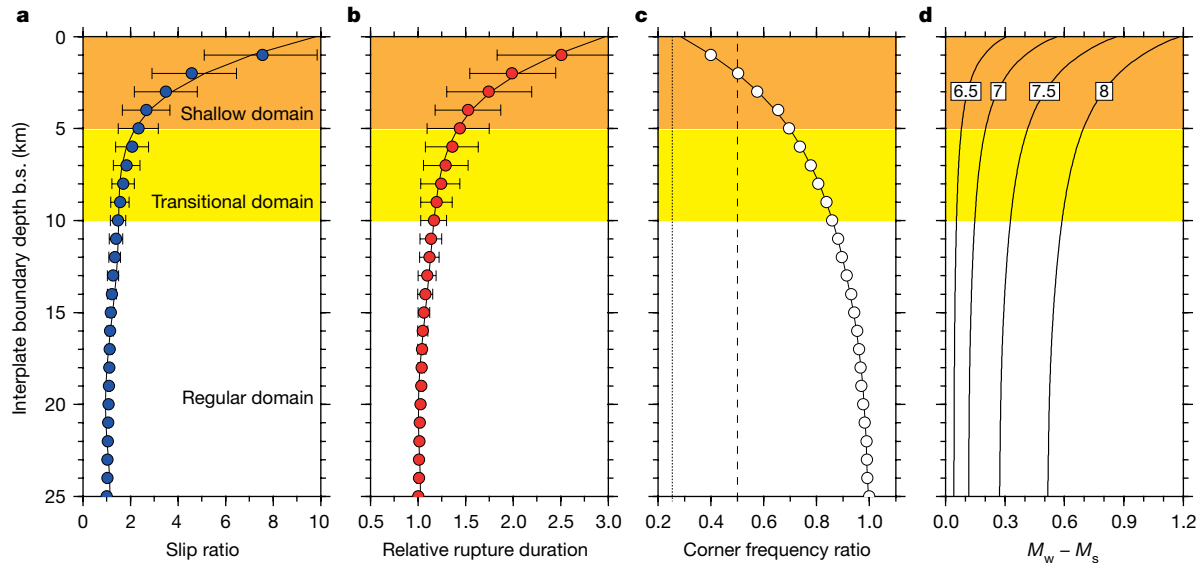
shown in Extended Data Fig. 1. Additional information and references are provided in Extended Data Table 1. Blue lines with arrowheads indicate correspondence between different depth domains in **a** and  $V_p$  distribution (**b**). **c**, Orange circles show average  $V_p$  as a function of  $z$ , obtained by averaging  $V_p(z)$  values in **b** within a 1-km-thick sliding window. **d**, Blue circles show  $\mu$  as a function of  $z$ , obtained from  $\rho(z)$  and  $V_s(z)$  in Extended Data Fig. 4. Data point values in **c** and **d** are fitted with a fourth-order polynomial regression (black lines). Error bars represent one standard deviation.

The depth trend of elastic properties within the three domains strongly conditions the predicted differences in rupture characteristics. To show this, we compare predicted ratios of amount of slip, rupture duration and corner frequency, as well as  $M_w - M_s$ , for earthquakes of equal magnitude and equal stress drop<sup>20</sup>, computed from classical self-similar source theory taking as a reference the values at 25 km depth (Fig. 3 and Methods). Our results show that, for all the source properties considered, relative changes are concentrated in the shallow domain. For a given earthquake magnitude and stress drop, the predicted amount of slip is about 5–10 times larger in the shallow domain than in the regular domain (Fig. 3a), whereas rupture duration is 2–3 times larger (Fig. 3b), and corner frequency ( $f_c$ ) 1–2 octaves lower (Fig. 3c). The  $f_c$  lowering implies that shallow earthquakes should be depleted in high frequencies. The high-frequency depletion produces a depth-dependent discrepancy between  $M_w$  and  $M_s$  because these two earthquake magnitudes are based on data at different frequencies (Extended Data Fig. 6b). The predicted  $M_w - M_s$  difference for a  $M_w$  7.5 event is 0.2–0.3 in the regular domain but increases to 0.6–0.8 in the shallow domain because of the decrease in  $f_c$  (Fig. 3d). Figure 4 presents a conceptual model summarizing all these predictions.

The obtained values agree with average trends of rupture properties of natural examples. One example is tsunami earthquakes, infrequent

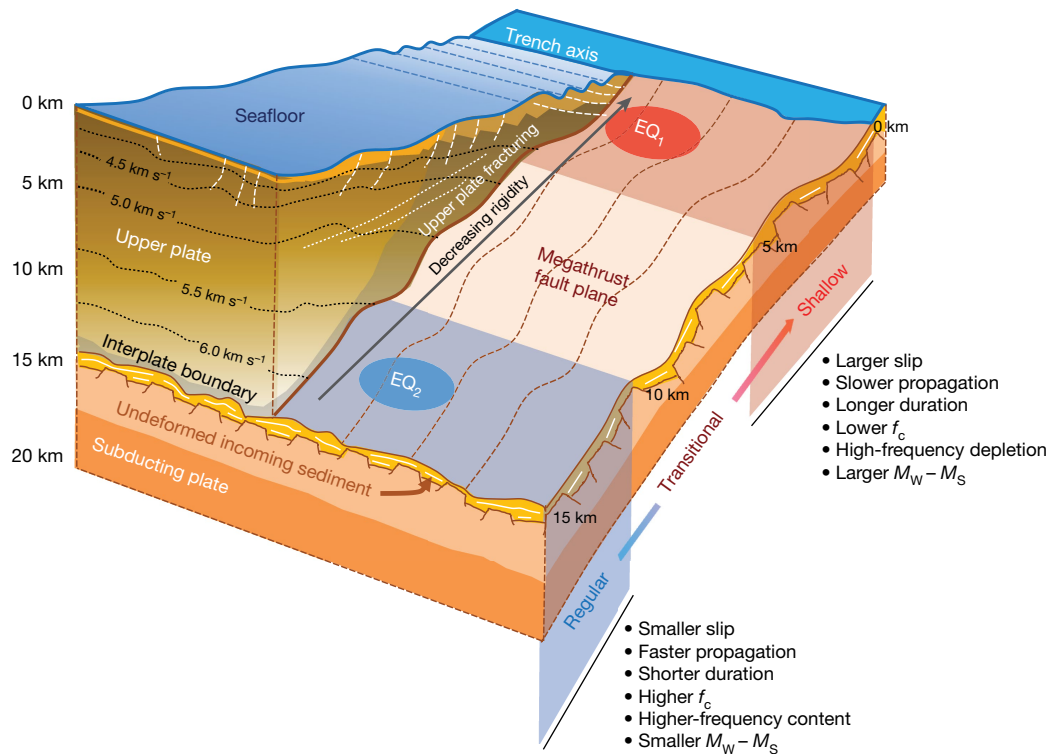
but well-documented events that rupture only the shallowest megathrust and generate anomalously large tsunamis for their magnitude<sup>11,21</sup> (Extended Data Fig. 1 and Extended Data Table 2). These events display all the characteristics of shallow ruptures, including long duration, high-frequency depletion inducing subdued seismic shaking, and large  $M_w - M_s$  discrepancy<sup>13,21,22</sup>. These characteristics, however, are not unique to tsunami earthquakes. Great earthquakes rupturing from deep in the seismogenic zone to close to the trench axis exhibit similar rupture properties in their shallow-depth portion<sup>1</sup>.

Studies of tsunami earthquakes based on seismological, geodetic and tsunami modelling support the idea that slip not only concentrated in the shallow domain, but it also increased upwards to peak near the trench axis<sup>13,23</sup>. Likewise, slip of some great tsunamigenic earthquakes appears to have peaked close to the trench, most clearly for the  $M_w$  9.1 Tohoku-Oki event in 2011, with maximum slip exceeding 50 m near the trench axis<sup>9,24</sup>, and for the  $M_w$  8.8 Maule event in 2010<sup>10</sup>. Current understanding attributes large shallow slip to the frictional properties of fault-rock materials<sup>25</sup>, or to local features such as near-trench slumps<sup>26</sup>, and to subducting relief<sup>27</sup>. However, the 4–5-fold decrease in  $\mu$  in the shallowest part of the megathrust (Fig. 2d) implies a 5–10-fold increase in slip relative to regular earthquakes of the same size (Fig. 3a). This trenchward slip increase is consistent with the large shallow slip



**Fig. 3 | Predicted earthquake rupture and energy release characteristics.** **a**, Blue circles show slip ratio ( $D_R$ ) for an earthquake of a given magnitude as a function of interplate boundary depth b.s. ( $z$ ).  $D_R$  is calculated taking as reference unit the slip at  $z = 25$  km. **b**, Red circles show relative mode III rupture duration ( $T_R$ ) for an earthquake of a given magnitude, as a function of  $z$ .  $T_R$  is calculated taking as reference the rupture duration per unit length at  $z = 25$  km. **c**, White circles show corner frequency ratio ( $f_R$ ) for an earthquake of a given magnitude as a function of  $z$ .  $f_R$  is calculated taking as reference unit the corner

frequency at  $z = 25$  km. The dashed and dot-dashed black lines show frequencies one and two octaves lower than the reference one, respectively. **d**, Black lines show the difference between  $M_w$  and  $M_s$  as a function of  $z$ , for earthquakes of  $M_w = 6.5, 7, 7.5$  and  $8$ . Orange, yellow and white rectangles in all panels indicate the depth extension of the shallow, transitional and regular domains referred to in the text. See details on the calculations in the main text. Data point values in **a–c** are fitted with a fourth-order polynomial regression (black lines). The error bars represent one standard deviation.



**Fig. 4 | Conceptual model of megathrust seismogenic zone domains.** Schematic diagram shows main geological and tectonic features of the upper and subducting plate based on geophysical data of Central and South America. Dotted black lines in the upper plate indicate isovelocity contours. Overriding plate faulting and fracturing increases towards the trench, thereby reducing  $V_p$ ,  $\rho$ ,  $V_s$  and rigidity ( $\mu$ ). Deformation and fracturing are concentrated above subducting sediment and interplate boundary. The diagram also shows differences of earthquake rupture and energy release characteristics for megathrust earthquakes occurring within the shallow (red) and regular (blue) domains discussed in main text. Red (blue) ellipses labelled EQ<sub>1</sub> (EQ<sub>2</sub>) are

rupture zones of the same size occurring within the shallow (regular) domains. Depth-dependent changes in elastic properties quantitatively explain that, compared with EQ<sub>2</sub>, EQ<sub>1</sub> should have propagation velocity up to 2–3 times slower, so 2–3 times longer duration; 5–10 times larger slip, so high tsunamigenic potential; 1–2 octaves lower corner frequency, so high-frequency depletion and subdued seismic shaking; and a 3–4 times larger  $M_w - M_s$  discrepancy, of up to 0.6–0.8 for a  $M_w 7.5$  earthquake (Fig. 3). The model predicts that the shallowest ruptures, and especially those reaching the trench axis, should show the largest differences with respect to regular events in all the analysed attributes, as is observed in natural examples.

required to generate large tsunamis by either tsunami earthquakes<sup>12,14</sup> or great earthquakes rupturing to the trench<sup>28</sup>. Specifically, a 5-fold reduction of  $\mu$  between regular and shallow domain depths was inferred to fit tsunami wave amplitudes of the  $M_s$  7.0–7.2 Nicaragua tsunami earthquake in 1992<sup>15</sup> (Extended Data Fig. 1 and Extended Data Table 2).

The slow rupture propagation and long duration compared with deeper events of the same magnitude is a key characteristic of tsunami earthquakes, to the point that they are often referred to as ‘slow tsunami earthquakes’<sup>12,13</sup>. Their average propagation velocity is about  $1\text{--}2\text{ km s}^{-1}$  (ref. 22), whereas for deeper events it is about  $3\text{ km s}^{-1}$ , in agreement with predicted propagation velocity differences between the shallow and regular domains (Extended Data Fig. 4c). The predicted increase of source duration also agrees with observations of normalized duration for  $M_w$  5.0–7.5 earthquakes in circum-Pacific subduction zones<sup>24</sup> (Extended Data Fig. 7a). These data show that the duration of earthquakes shallower than about 10 km depth, which include six tsunami earthquakes, is 2–3 times that of deeper events, as predicted by our model (Fig. 3b and Extended Data Fig. 7b). Smaller-magnitude events occurring within the rupture areas of the  $M_w$  9.1 2004 Sumatra-Andaman<sup>29</sup> and the  $M_w$  9.1 2011 Tohoku-Oki<sup>30</sup> earthquakes also show longer normalized duration in the near-trench zone.

Another characteristic of tsunami earthquakes is a high-frequency deficit relative to regular events of equal magnitude<sup>13</sup>. The resulting ground shaking is weaker, and tsunami hazard based on human perception is therefore underestimated. This was the case of the 1992 Nicaragua earthquake, where mild shaking caused little damage, and the tsunami hit the coast unexpectedly. But this feature is not unique to tsunami earthquakes. Seismological data of recent great tsunamigenic earthquakes support a pattern of two distinct rupture modes for the  $M_w$  9.1 2004 Sumatra-Andaman<sup>1</sup>, the  $M_w$  9.1 2011 Tohoku-Oki<sup>31</sup>, the  $M_w$  8.8 2010 Maule<sup>32</sup> and the  $M_w$  8.3 2015 Illapel<sup>33</sup> events (Extended Data Fig. 1 and Extended Data Table 2). Those earthquakes initiated deep into the seismogenic zone with rupture radiating high-frequency energy and producing strong shaking, followed by shallow rupture with lower-frequency content that generated large seafloor deformation and originated the tsunamis. The trend of higher-frequency content in the regular domain (Fig. 3c and Extended Data Fig. 6a) due to the spectral amplitude decay (Extended Data Fig. 6b) can also be explained by the depth-dependent overriding rock properties, without calling for a hypothetical depth-dependent stress drop trend that is barely supported by seismological data<sup>20</sup>.

Owing to high-frequency depletion, the initial magnitude estimation of the 1992 Nicaragua earthquake was  $M_s$  6.8 (later corrected to 7.0–7.2), too low for a tsunami alert to be issued. However, the  $M_w$  7.6–7.8 calculated after more detailed data analysis, if available earlier, would have prompted the alert. On average, tsunami earthquakes have  $|M_w| \approx 7.6$  and  $|M_w - M_s| \approx 0.65$  (Extended Data Table 2), so that they have larger  $M_w - M_s$  differences than regular earthquakes of the same magnitude. These  $M_w - M_s$  discrepancies are difficult to explain for earthquakes of magnitude  $M_w$  7.6 rupturing the regular domain, where  $M_w - M_s$  should be less than about 0.3 (Fig. 3d). However, the depth-dependent elastic properties imply that the  $M_w - M_s$  discrepancy for earthquakes of this magnitude can increase to 0.6–0.8 when rupture concentrates in the shallow domain (Fig. 3d), in agreement with observations.

Although shallow megathrust earthquake ruptures are infrequent, their slip distribution peaking near the trench makes them particularly hazardous. Extended Data Fig. 8 shows that a  $M_w$  7 earthquake rupturing the regular domain has the same spectral amplitude at 20 s as a  $M_w$  8 event rupturing the shallow domain, and thus the same  $M_s$ , if depth-dependent changes of elastic properties are taken into account. The associated tsunami hazard of these two events is radically different, but it cannot be forecast based on  $M_s$ . Proper magnitude and tsunami hazard evaluation require incorporating focal depth information and the local  $V_p(z)$ . In the interim lack of local velocity models, tsunami forecast can be improved using the global trends obtained here (Fig. 2).

In summary, interplate fault mechanics may play a role controlling different aspects of the seismic cycle but do not seem to be required to explain the overall depth-dependent trend of the source properties considered here. We show quantitatively that the observed characteristics of both shallow and regular earthquakes reflect the elastic properties of the rock volume undergoing dynamic stress transfer. However, our model uses average physical properties to explain global trends of source characteristics rather than individual examples. The observed variability of physical properties across different systems (Fig. 2b) implies that proper analysis of particular seismic events would require determination of the elastic properties throughout their specific rupture zone. These properties should be incorporated into numerical models to allow evaluation of the potential effect of complex fault mechanics on rupture characteristics.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1784-0>.

- Lay, T. et al. Depth-varying rupture properties of subduction zone megathrust faults. *J. Geophys. Res.* **117**, B04311 (2012).
- Lay, T. & Bilek, S. L. in *The Seismogenic Zone of Subduction Thrust Faults* (eds. Dixon, T. & Moore, C.) 476–511 (Columbia Univ. Press, 2007).
- Bilek, S. L. & Lay, T. Subduction zone megathrust earthquakes. *Geosphere* **14**, 1468–1500 (2018).
- Tobin, H. J. & Saffer, D. M. Elevated fluid pressure and extreme mechanical weakness of a plate boundary thrust, Nankai Trough subduction zone. *Geology* **37**, 679–682 (2009).
- Noda, H. & Lapusta, N. Stable creeping fault segments can become destructive as a result of dynamic weakening. *Nature* **493**, 518–521 (2013).
- Huang, Y., Meng, L. & Ampuero, J.-P. A dynamic model of the frequency-dependent rupture process of the 2011 Tohoku-Oki earthquake. *Earth Planets Space* **64**, 1061–1066 (2012).
- Ikari, M. J., Kameda, J., Saffer, D. M. & Kopf, A. J. Strength characteristics of Japan Trench borehole samples in the high-slip region of the 2011 Tohoku-oki earthquake. *Earth Planet. Sci. Lett.* **412**, 35–41 (2015).
- Scholz, C. Earthquakes and friction laws. *Nature* **391**, 37–42 (1998).
- Fujiwara, T. et al. The 2011 Tohoku-Oki earthquake: displacement reaching the trench axis. *Science* **334**, 1240 (2011).
- Maksymowicz, A. et al. Coseismic seafloor deformation in the trench region during the  $M_w$  8.8 Maule megathrust earthquake. *Sci. Rep.* **7**, 45918 (2017).
- Kanamori, H. Mechanism of tsunami earthquakes. *Phys. Earth Planet. Inter.* **6**, 346–359 (1972).
- Kanamori, H. & Kikuchi, M. The 1992 Nicaragua earthquake: a slow tsunami earthquake associated with subducted sediments. *Nature* **361**, 714–716 (1993).
- Polet, J. & Kanamori, H. Shallow subduction zone earthquakes and their tsunamigenic potential. *Geophys. J. Int.* **142**, 684–702 (2000).
- Satake, K. & Tanioka, Y. Sources of tsunami and tsunamigenic earthquakes in subduction zones. *Pure Appl. Geophys.* **154**, 467–483 (1999).
- Geist, E. L. & Bilek, S. L. Effect of depth-dependent shear modulus on tsunami generation along subduction zones. *Geophys. Res. Lett.* **28**, 1315–1318 (2001).
- Bilek, S. L. & Lay, T. Rigidity variations with depth along interplate megathrust faults in subduction zones. *Nature* **400**, 443–446 (1999).
- von Huene, R., Klaeschen, D., Cropp, B. & Miller, J. Tectonic structure across the accretionary and erosional parts of the Japan Trench margin. *J. Geophys. Res.* **99**, 22, 349–22,361 (1994).
- Ranero, C. R. & von Huene, R. Subduction erosion along the Middle America convergent margin. *Nature* **404**, 748–752 (2000).
- Brocher, T. M. Empirical relations between elastic wavespeeds and density in the Earth's crust. *Bull. Seismol. Soc. Am.* **95**, 2081–2092 (2005).
- Allmann, B. P. & Shearer, P. M. Global variations of stress drop for moderate to large earthquakes. *J. Geophys. Res.* **114**, B01310 (2009).
- Bilek, S. L. & Lay, T. Tsunami earthquakes possibly widespread manifestations of frictional conditional stability. *Geophys. Res. Lett.* **29**, 18–18-4 (2002).
- Pelayo, A. M. & Wiens, D. A. Tsunami earthquakes; slow thrust-faulting events in the accretionary wedge. *J. Geophys. Res.* **97**, 15321–15337 (1992).
- Newman, A. V., Hayes, G., Wei, Y. & Convers, J. The 25 October 2010 Mentawai tsunami earthquake, from real-time discriminants, finite-fault rupture, and tsunami excitation. *Geophys. Res. Lett.* **38**, L05302 (2011).
- Sun, T., Wang, K., Fujiwara, T., Kodaira, S. & He, J. Large fault slip peaking at trench in the 2011 Tohoku-Oki earthquake. *Nat. Commun.* **8**, 14044 (2017).
- Hirono, T. et al. Near-trench slip potential of megaquakes evaluated from fault properties and conditions. *Sci. Rep.* **6**, 28184 (2016).
- Okal, A. & Newman, A. V. Tsunami earthquakes: the quest for a regional signal. *Phys. Earth Planet. Inter.* **124**, 45–70 (2001).
- Abercrombie, R. E., Antolik, M., Felzer, K. & Ekström, G. The 1994 Java tsunami earthquake: slip over a subducting seamount. *J. Geophys. Res.* **106**, 6595–6607 (2001).

28. Murphy, S. et al. Shallow slip amplification and enhanced tsunami hazard unravelled by dynamic simulations of mega-thrust earthquakes. *Sci. Rep.* **6**, 35007 (2016).
29. Bilek, S. L. Using earthquake source durations along the Sumatra-Andaman subduction system to examine fault-zone variations. *Bull. Seismol. Soc. Am.* **97**, S62–S70 (2007).
30. Bilek, S. L., DeShon, H. R. & Engdahl, E. R. Spatial variations in earthquake source characteristics within the 2011  $M_w = 9.0$  Tohoku, Japan rupture zone. *Geophys. Res. Lett.* **39**, L09304 (2012).
31. Wang, D. & Mori, J. Frequency-dependent energy radiation and fault coupling for the 2010  $M_w 8.8$  Maule, Chile, and 2011  $M_w 9.0$  Tohoku, Japan, earthquakes. *Geophys. Res. Lett.* **38**, L22308 (2011).
32. Koper, K. D., Hutko, A. R., Lay, T. & Sufri, O. Imaging short-period seismic radiation from the 27 February 2010 Chile ( $M_w 8.8$ ) earthquake by back-projection of P, PP, and PKIKP waves. *J. Geophys. Res.* **117**, B02308 (2012).
33. Melgar, D. et al. Slip segmentation and slow rupture to the trench during the 2015,  $M_w 8.3$  Illapel, Chile earthquake. *Geophys. Res. Lett.* **43**, 961–966 (2016).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

### Joint reflection and refraction travel-time modelling of wide-angle seismic data

The methods used to obtain the 2D  $V_p$  models included in our compilation (Extended Data Table 1) include both forward<sup>34</sup> and inverse<sup>35–37</sup> techniques. All the selected profiles (Extended Data Fig. 1) share two key aspects. First, they are travel-time-fitting techniques based on ray-tracing approaches, and second, they include seismic phases reflected at the interplate boundary. The joint modelling of first arrival (that is, refracted waves within the overriding and subducting plates) and interplate reflection travel-times allow mapping of not only the  $V_p$  structure but also the location and geometry of the interplate boundary, from the trench to depths ranging between about 15 km and 30 km, depending on data quality and experiment set-up. When 2D  $V_p$  models and multichannel seismic data (Fig. 1) are spatially coincident, they can be combined to improve the geological interpretation of seismic velocities (Extended Data Fig. 2). Monte-Carlo-type statistical analysis of several profiles, with multiple inversions using different initial models and assuming realistic travel-time picking errors, provides  $V_p$  uncertainty. Above the interplate boundary, this is typically 0.05–0.1 km s<sup>-1</sup> at the shallowest megathrust sector and 0.2–0.3 km s<sup>-1</sup> at about 25 km depth<sup>38–41</sup>.

### Resolution of travel-time-based seismic modelling versus wavelength of the stress wavefield

Using seismic velocity models to infer earthquake rupture-related properties implicitly assumes that rupture propagation is affected by the properties of a rock volume that can be resolved by  $V_p$  models. Rupture initiation depends on the stress distribution surrounding the crack tip, and the subsequent rupture propagation and material deformation reflect the dynamic stress transfer around the crack tip<sup>42</sup>. Rupture propagation velocity is limited by the speed at which stresses can propagate through the material (that is,  $V_s$  for mode III rupture)<sup>43</sup>. Additionally, near-field ground motion recordings of large subduction earthquakes consistently display a peak frequency  $f_{sw}$  of about 1–4 Hz (refs.<sup>44,45</sup>). This implies that the stress transfer, whose limiting propagation velocity along the megathrust varies with depth as indicated in Extended Data Fig. 4c, has an associated wavelength  $\lambda_{sw}(z) = V_s(z)/f_{sw}$ , ranging from about 0.5–1.5 km near the surface to about 1.5–4.0 km at 25 km depth (blue-lilac polygon in Extended Data Fig. 3).

Modern wide-angle seismic data can resolve  $V_p$  of the rock body equivalent to the wavelength of the stress wavefield. For ray-based, travel-time-fitting methods such as the ones used in this study (Extended Data Table 1), model resolution is inferred to be limited to the width of the first Fresnel zone,  $R_F = \sqrt{(zV_p/2f_s)}$ , where  $z$  is the imaged target depth,  $V_p$  is P-wave velocity and  $f_s$  is the peak frequency of the seismic source. Taking the  $V_p(z)$  values in Extended Data Fig. 2c and  $f_s = 8–12$  Hz, which is the typical frequency content of records, we obtain  $R_F(z)$  ranging from about 0.3–0.4 km near the trench to about 2.5–3.5 km at 25 km depth (light red area in Extended Data Fig. 3). The comparable size of the region resolved by travel-time-based velocity models at all depths and the wavelength of the seismic wavefield associated to rupture propagation (Extended Data Fig. 3), supports the idea that the modelled  $V_p(z)$  in Fig. 2, as well as other  $V_p$ -derived properties (Fig. 2d and Extended Data Fig. 4), represents the physical properties influencing the dynamic stress transfer associated to the propagating seismic rupture.

### Estimation of $V_p(z)$ above the interplate boundary

To obtain the  $V_p$  values as a function of interplate boundary depth below the seafloor (Fig. 2), we first digitized  $V_p$  just above the interplate boundary, interplate boundary depth, and seafloor depth/land topography, along the 48 wide-angle seismic profiles (Extended Data Table 1). Second, we interpolated  $V_p$ , interplate boundary depth below

sea surface ( $z_i$ ), and seafloor depth/land topography ( $z_o$ ) at constant  $x$  intervals (2 km) along each profile by applying Akima splines to obtain  $V_p$  as a function of upper plate thickness,  $z = z_i - z_o$ , using Generic Mapping Tools (GMT)<sup>46</sup>. For simplicity, this value is referred to as ‘interplate boundary depth b.s.’ throughout the manuscript and in the figures. Third, we interpolated  $V_p$  at constant  $z$  intervals (1 km) along each profile, again using GMT. Fourth, for each  $z$  between 1 km and 25 km, we calculated the average  $V_p$  value of all profiles and its corresponding standard deviation. Finally, we used GMT to calculate a fourth-order polynomial regression fit of the  $V_p(z)$  values.

### Derivation of rock properties from $V_p(z)$

The  $V_p(z)$  values shown in Fig. 2c are used as a reference to calculate the rest of physical properties presented throughout the manuscript as a function of  $z$ . The shear-wave modulus, or rigidity,  $\mu = \rho V_s^2$  (Fig. 2d), is obtained by applying first empirically based relations proposed in ref.<sup>19</sup> to estimate  $\rho$  (Extended Data Fig. 4b) and  $V_s$  (Extended Data Fig. 4c) from  $V_p$ . For density, the relation is  $\rho = 1.6612V_p - 0.4721V_p^2 + 0.0671V_p^3 - 0.0043V_p^4 + 0.000106V_p^5$ , whereas for shear velocity, it is  $V_s = 0.7858 - 1.2344V_p + 0.7949V_p^2 - 0.1238V_p^3 + 0.0064V_p^4$ . Both relationships are based on a compilation of  $V_p$ ,  $V_s$  and  $\rho$  measurements for a wide variety of Earth crustal rocks, obtained from wireline borehole logs, vertical seismic profiles, laboratory measurements and seismic tomography models.

### Derivation of depth-dependent earthquake rupture characteristics

**Amount of slip.** The seismic moment released during co-seismic rupture is  $M_0 = \int_S \mu D ds \approx \bar{\mu} \bar{D} S$ , where  $D$  is co-seismic slip,  $\mu$  is rigidity,  $ds$  is a surface element, and the integral runs over the whole rupture area,  $S$ . If we take average values over  $S$  ( $\bar{D}$  and  $\bar{\mu}$ ),  $M_0$  is approximated by the right-hand side.  $M_0$  and  $S$  can also be estimated from waveform data and aftershock location, respectively, but  $\mu$  and  $D$  are generally unknown, so they have a trade-off in the calculations. To show the effect of  $\mu(z)$  for events of equal  $M_0$  and  $S$ , we compare differences in the amount of slip as a function of depth as a slip ratio  $D_R(z)$ . Using as a unit reference the amount of slip at the regular domain depth of 25 km,  $D_R(z)$  can be calculated as  $D_R(z) = \frac{D(z)}{D^*} = \frac{\mu^*}{\mu(z)}$ , where  $D^*$  and  $\mu^*$  are the amount of slip and rigidity at the reference depth of 25 km, and  $D(z)$ ,  $\mu(z)$  are at depth  $z$ . Figure 3a displays  $D_R(z)$  calculated by taking  $\mu(z)$  from the global compilation (Fig. 2d).

**Rupture duration.** Given that stresses must accumulate at the crack tip for spontaneous propagation, rupture velocity ( $u$ ) is limited by the velocity at which stresses are transmitted throughout the material. For mode III cracks, in which a shear stress acts parallel to the plane of the crack and parallel to the crack front, as in megathrust fault ruptures, the theoretical limiting velocity is  $V_s$ . Field data indicate that  $u$  is actually 70–90% of  $V_s$  at the depth of the largest slip. Thus, the observed  $V_s(z)$  trend above the megathrust (Extended Data Fig. 4c) implies that  $u$  should be significantly lower in the shallow domain than in the regular domain. To illustrate this effect, we calculated the normalized source duration for a unit rupture length at different depths,  $T_R(z)$ , using  $u(z)$  in Extended Data Fig. 4c and taking as a reference the rupture duration at the regular domain depth of 25 km, as  $T_R(z) = \frac{T(z)}{T^*} = \frac{u^*}{u(z)} = \frac{V_s^*}{V_s(z)}$ , where  $T^*$ ,  $u^*$  and  $V_s^*$  are rupture duration per unit length, rupture velocity and shear velocity at the reference depth of 25 km, and  $T(z)$ ,  $u(z)$ ,  $V_s(z)$  are at depth  $z$ .

**Corner frequency.** The  $V_s(z)$  distribution also influences the frequency content of the energy released during an earthquake. This can be estimated from the spectral shape of the moment-rate spectrum,  $\dot{M}(f)$ , which is calculated based on waveform records. The reference spectra to compare with values obtained from observational data are



$\dot{M}(f) = \frac{M_0 f_c^n}{f^n + f_c^n}$ , where  $f_c$  is the corner frequency, which marks the frequency at which the spectrum starts to decay, and  $n$  is the slope of the spectral decay. A value of  $n=2$  is typically used as a reference, as it fits the observed decay in most cases, whereas  $f_c = cV_s \left( \frac{\Delta\sigma}{M_0} \right)^{\frac{1}{3}}$ , where  $c=0.49$  is a dimensionless constant, and  $\Delta\sigma$  is stress drop. Thus,  $V_s$  of the region enclosing the propagating rupture front determines  $f_c$  and therefore the spectral shape. In contrast to  $V_p$  and  $V_s$ , there is no clear evidence indicating a systematic, universal depth-dependence of  $\Delta\sigma$  in subduction zones<sup>20</sup>. Therefore, we assume that it is depth-independent, and we isolate the  $V_s(z)$  effect on the corner frequency ratio,  $f_R(z)$ , for events of equal  $M_0$ , using  $f_c(V_s)$ . Taking as a reference the  $f_c$  and  $V_s$  at the regular domain depth of 25 km, we have  $f_R(z) = \frac{f_c(z)}{f_c} = \frac{V_s(z)}{V_s} = T_R(z)^{-1}$ , where  $f_c^*$  and  $V_s^*$  are the corner frequency and shear-wave velocity at the reference depth of 25 km, and  $f_c(z)$ ,  $V_s(z)$  and  $T_R(z)$  are at depth  $z$ .

**Moment magnitude versus surface wave magnitude.** Another consequence of the  $V_s$ -dependent high-frequency depletion is a depth-dependent discrepancy between the earthquake moment magnitude  $M_w$ , estimated from long-period waves (approximately 250 s), and its surface wave magnitude  $M_s$ , estimated at shorter periods (about 20 s) (Extended Data Fig. 6b). This effect is illustrated in Fig. 3d, where each curve represents the difference between  $M_w$  and  $M_s$  calculated from the moment-rate spectrum as a function of depth  $\dot{M}(f) = \frac{M_0 f_c^n}{f^n + f_c^n}$  for four different  $M_w$ . Given that  $f_c$  is anticorrelated with  $M_0$  according to  $f_c = cV_s \left( \frac{\Delta\sigma}{M_0} \right)^{\frac{1}{3}}$ ,  $M_s$  tends to saturate for large magnitudes, so that the  $M_w - M_s$  discrepancy increases with increasing  $M_w$ . However, Fig. 3d shows that the discrepancy for any magnitude is also depth-dependent.

## Data availability

Source data for Figs. 2 and 3 and for Extended Data Figs. 4–8 are provided with the paper. The digitized values of P-wave seismic velocity above interplate boundary versus depth and seafloor depth along the 48 wide-angle seismic profiles used here are available at the public research data repository figshare (<https://doi.org/10.6084/m9.figshare.9729302.v1>).

## Code availability

The scripts necessary to process the data and reproduce the main results and figures presented in this work are available at the public research data repository figshare (<https://doi.org/10.6084/m9.figshare.9729302.v1>).

34. Zelt, C. A. & Smith, R. B. Seismic travel time inversion for 2-D crustal velocity structure. *Geophys. J. Int.* **108**, 16–34 (1992).
35. Van Avendonk, H. J. A., Harding, A. J., Orcutt, J. A. & McClain, J. S. A two-dimensional tomographic study of the Clipperton transform fault. *J. Geophys. Res.* **103**, 17885–17899 (1998).
36. Korenaga, J. et al. Crustal structure of the southeast Greenland margin from joint refraction and reflection seismic tomography. *J. Geophys. Res.* **105**, 21591–21614 (2000).
37. Hobro, J. W. D., Singh, S. C. & Minshull, T. A. Three-dimensional tomographic inversion of combined reflection and refraction seismic traveltimes. *Geophys. J. Int.* **152**, 79–93 (2003).
38. Contreras-Reyes, E., Grevemeyer, I., Flueh, E. R. & Reichert, C. Upper lithospheric structure of the subduction zone offshore of southern Arauco peninsula, Chile, at 38° S. *J. Geophys. Res.* **113**, B07303 (2008).
39. Contreras-Reyes, E. et al. Structure and tectonics of the central Chilean margin (31°–33° S): implications for subduction erosion and shallow crustal seismicity. *Geophys. J. Int.* **203**, 776–791 (2015).
40. Sallarès, V. & Ranero, C. R. Structure and tectonics of the erosional convergent margin off Antofagasta, north Chile (23.30° S). *J. Geophys. Res.* **110**, B06101 (2005).
41. Sallarès, V. et al. Overriding plate structure of the Nicaragua convergent margin: relationship to the seismogenic zone of the 1992 tsunami earthquake. *Geochem. Geophys. Geosyst.* **14**, 3436–3461 (2013).
42. Svetlizky, I. & Fineberg, J. Classical shear cracks drive the onset of dry frictional motion. *Nat. Phys.* **509**, 205–208 (2014).
43. Freund, L. B. *Dynamic Fracture Mechanics* (Cambridge Univ. Press, 1998).
44. Ambraseys, N. N. & Douglas, J. Near-field horizontal and vertical earthquake ground motions. *Soil. Dyn. Earthquake Eng.* **23**, 1–18 (2003).
45. Atkinson, G. M. & Boore, D. M. Empirical ground-motion relations for subduction-zone earthquakes and their application to Cascadia and other regions. *Bull. Seismol. Soc. Am.* **93**, 1703–1729 (2003).
46. Wessel, P., Smith, W. H. F., Scharroo, R., Luis, J. & Wobbe, F. Generic Mapping Tools: improved version released. *Eos* **94**, 409–410 (2013).
47. IOC, IHO and BODC, Centenary Edition of the GEBCO Digital Atlas. [https://www.gebco.net/data\\_and\\_products/gridded\\_bathymetry\\_data](https://www.gebco.net/data_and_products/gridded_bathymetry_data) (2003).
48. Martínez-Loriente, S. et al. Influence of incoming plate relief on upper plate structure and on earthquake nucleation: the case of Southern Costa Rica. *Tectonics* (in the press).
49. Agudelo, W., Ribodetti, A., Collot, J.-Y. & Operto, S. Joint inversion of multichannel seismic reflection and wide-angle seismic data: Improved imaging and refined velocity model of the crustal structure of the north Ecuador–south Colombia convergent margin. *J. Geophys. Res.* **114**, B02306 (2009).
50. Contreras-Reyes, E., Becerra, J., Kopp, H., Reichert, C. & Díaz-Naveas, J. Seismic structure of the north-central Chilean convergent margin: Subduction erosion of a paleomagmatic arc. *Geophys. Res. Lett.* **41**, 1523–1529 (2014).
51. Moscoso, E. et al. Revealing the deep structure and rupture plane of the 2010 Maule, Chile earthquake ( $M_w = 8.8$ ) using wide angle seismic data. *Earth Planet. Sci. Lett.* **307**, 147–155 (2011).
52. Scherwath, M. et al. Deep lithospheric structures along the southern central Chile margin from wide-angle P-wave modelling. *Geophys. J. Int.* **179**, 579–600 (2009).
53. Contreras-Reyes, E. et al. Deep seismic structure of the Tonga subduction zone: implications for mantle hydration, tectonic erosion, and arc magmatism. *J. Geophys. Res.* **116**, B10103 (2011).
54. Klingelhoefer, F. et al. Limits of the seismogenic zone in the epicentral region of the 26 December 2004 great Sumatra-Andaman earthquake: Results from seismic refraction and wide-angle reflection surveys and thermal modeling. *J. Geophys. Res.* **115**, B01304 (2010).
55. Arai, R. et al. Structure of the tsunamigenic plate boundary and low-frequency earthquakes in the southern Ryukyu Trench. *Nat. Commun.* **7**, 12255 (2016).
56. Kodaira, S., Takahashi, N., Nakanishi, A., Miura, S. & Kaneda, Y. Subduction seamount imaged in the rupture zone of the 1946 Nankaido earthquake. *Science* **289**, 104–106 (2000).
57. Nakamura, Y. et al. Seismic imaging and velocity structure around the JFAST drill site in the Japan Trench: low  $V_p$ , high  $V_p/V_s$  in the transparent frontal prism. *Earth Planets Space* **66**, 121 (2014).
58. Horning, G. et al. 2-D tomographic model of the Juan de Fuca plate from accretion at axial seamount to subduction at the Cascadia margin from an active source ocean bottom seismometer survey. *J. Geophys. Res.* **121**, 5859–5879 (2016).
59. Krabbenhöft, A., Bialas, J., Kopp, H., Kukowski, N. & Hübner, C. Crustal structure of the Peruvian continental margin from wide-angle seismic studies. *Geophys. J. Int.* **159**, 749–764 (2004).
60. Hampel, A., Kukowski, N., Bialas, J., Huebscher, C. & Heinbockel, R. Ridge subduction at an erosive margin: the collision zone of the Nazca Ridge in southern Peru. *J. Geophys. Res.* **109**, B02101 (2004).
61. Shulgin, A. et al. Structural architecture of oceanic plateau subduction offshore Eastern Java and the potential implications for geohazards. *Geophys. J. Int.* **184**, 12–28 (2011).
62. Bassett, D. et al. Three dimensional velocity structure of the northern Hikurangi margin, Raukumara, New Zealand: implications for the growth of continental crust by subduction erosion and tectonic underplating. *Geochem. Geophys. Geosyst.* **11**, Q10013 (2010).
63. Miura, S. et al. Seismological structure and implications of collision between the Ontong Java Plateau and Solomon Island Arc from ocean bottom seismometer-airgun data. *Tectonophysics* **389**, 191–220 (2004).
64. Takahashi, N., Suyehiro, K. & Shinohara, M. Implications from the seismic crustal structure of the northern Izu–Bonin arc. *Isl. Arc* **7**, 383–394 (1998).
65. Walther, C. H. E., Flueh, E. R., Ranero, C. R., von Huene, R. & Strauch, W. Crustal structure across the Pacific margin of Nicaragua: evidence for ophiolitic basement and a shallow mantle sliver. *Geophys. J. Int.* **141**, 759–777 (2000).
66. Sallarès, V., Dañoibeitia, J. J. & Flueh, E. Lithospheric structure of the Costa Rican Isthmus: effects of subduction zone magmatism on an oceanic plateau. *J. Geophys. Res.* **106**, 621–643 (2001).
67. Bassett, D. et al. Crustal structure of the Kermadec arc from MANGO seismic refraction profiles. *J. Geophys. Res.* **121**, 7514–7546 (2016).
68. Klingelhoefer, F. et al. P-wave velocity structure of the southern Ryukyu margin east of Taiwan: results from the ACTS wide-angle seismic experiment. *Tectonophysics* **578**, 50–62 (2012).
69. Kopp, H., Klaeschen, D., Flueh, E. R., Bialas, J. & Reichert, C. Crustal structure of the Java margin from seismic wide-angle and multichannel reflection data. *J. Geophys. Res.* **107**, ETG 1–1–ETG 1-24 (2002).
70. Planert, L. et al. Lower plate structure and upper plate deformational segmentation at the Sunda–Banda arc transition, Indonesia. *J. Geophys. Res.* **115**, B08107 (2010).
71. Ye, S., Flueh, E. R., Klaeschen, D. & von Huene, R. Crustal structure along the EDGE transect beneath the Kodiak shelf off Alaska derived from OBH seismic refraction data. *Geophys. J. Int.* **130**, 283–302 (1997).
72. Nakanishi, A. et al. Crustal evolution of the southwestern Kuril Arc, Hokkaido Japan, deduced from seismic velocity and geochemical structure. *Tectonophysics* **472**, 105–123 (2009).
73. Nakanishi, A. et al. Crustal structure across the coseismic rupture zone of the 1944 Tonankai earthquake, the central Nankai Trough seismogenic zone. *J. Geophys. Res.* **107**, EPM 2–1–EPM 2-21 (2002).
74. Kodaira, S. et al. Western Nankai Trough seismogenic zone: results from a wide-angle ocean bottom seismic survey. *J. Geophys. Res.* **105**, 5887–5905 (2000).
75. Singh, S. C. et al. Seismic evidence of bending and unbending of subducting oceanic crust and the presence of mantle megathrust in the 2004 Great Sumatra earthquake rupture zone. *Earth Planet. Sci. Lett.* **321–322**, 166–176 (2012).

76. Kopp, H. et al. Deep structure of the central Lesser Antilles Island Arc: relevance for the formation of continental crust. *Earth Planet. Sci. Lett.* **304**, 121–134 (2011).
77. Zhu, J. et al. Crustal structure of the central Costa Rica subduction zone: implications for basal erosion from seismic wide-angle data. *Geophys. J. Int.* **178**, 1112–1131 (2009).
78. Begovic, S., Ranero, C. R., Sallarès, V. & Grevemeyer, I. 2D velocity and interplate geometry model of the North Chile margin from joint refraction and wide-angle reflection travel time inversion. In *Subduction Interface Processes (SIP) Int. Conf.* (2017).
79. Graindorge, D., Calahorrano, A., Charvis, P., Collot, J.-Y. & Bethoux, N. Deep structures of the Ecuador convergent margin and the Carnegie Ridge, possible consequence on great earthquakes recurrence interval. *Geophys. Res. Lett.* **31**, L04603 (2004).
80. Gailler, A., Charvis, P. & Flueh, E. R. Segmentation of the Nazca and South American plates along the Ecuador subduction zone from wide angle seismic profiles. *Earth Planet. Sci. Lett.* **260**, 444–464 (2007).
81. Krabbenhoef, A., von Huene, R., Klaeschen, D. & Miller, J. J. Subduction-related structure in the  $M_w$  9.2, 1964 megathrust rupture area offshore Kodiak Island, Alaska. In *AGU Fall Meet.* <https://ui.adsabs.harvard.edu/abs/2016AGUFM.T11D2641K/abstract> (2016).
82. Miura, S. et al. Structural characteristics off Miyagi forearc region, the Japan Trench seismogenic zone, deduced from wide-angle reflection and refraction study. *Tectonophysics* **407**, 165–188 (2005).
83. Nishizawa, A. et al. Variations in seismic velocity distribution along the Ryukyu (Nansei-Shoto) Trench subduction zone at the northwestern end of the Philippine Sea plate. *Earth Planets Space* <https://doi.org/10.1186/s40623-017-0674-7> (2017).
84. Barrientos, S. E. & Ward, S. N. The 1960 Chile earthquake: inversion for slip distribution from surface deformation. *Geophys. J. Int.* **103**, 589–598 (1990).
85. Kanamori, H. The Alaska Earthquake of 1964: radiation of long-period surface waves and source mechanism. *J. Geophys. Res.* **75**, 5029–5040 (1970).
86. Lay, T. et al. The great Sumatra–Andaman earthquake of 26 December 2004. *Science* **308**, 1127–1133 (2005).
87. Koketsu, K. et al. A unified source model for the 2011 Tohoku earthquake. *Earth Planet. Sci. Lett.* **310**, 480–487 (2011).
88. Delouis, B., Nocquet, J.-M. & Vallée, M. Slip distribution of the February 27, 2010  $M_w$  = 8.8 Maule Earthquake, central Chile, from static and high-rate GPS, InSAR, and broadband teleseismic data. *Geophys. Res. Lett.* **37**, L17305 (2010).
89. Wu, F. T. & Kanamori, H. Source mechanism of February 4, 1965, Rat Island earthquake. *J. Geophys. Res.* **78**, 6082–6092 (1973).
90. Ihmlé, P. F., Gómez, J.-M., Heinrich, Ph. & Guibourg, S. The 1996 Peru tsunamigenic earthquake: broadband source process. *Geophys. Res. Lett.* **25**, 2691–2694 (1998).
91. Bell, R., Holden, C., Power, W., Wang, X. & Downes, G. Hikurangi margin tsunami earthquake generated by slow seismic rupture over a subducted seamount. *Earth Planet. Sci. Lett.* **397**, 1–9 (2014).
92. Newman, A. V. et al. The energetic 2010  $M_w$  7.1 Solomon Island tsunami earthquake. *Geophys. J. Int.* **186**, 775–781 (2011).
93. Ammon, C. J., Kanamori, H., Lay, T. & Velasco, A. A. The 17 July 2006 Java tsunami earthquake. *Geophys. Res. Lett.* **33**, L24308 (2006).
94. Tanioka, T. & Satake, K. Fault parameters of the 1896 Sanriku tsunami earthquake estimated from tsunami numerical modeling. *Geophys. Res. Lett.* **23**, 1549–1552 (1996).
95. Johnson, J. M. & Satake, K. Estimation of seismic moment and slip distribution of the April 1, 1946, Aleutian tsunami earthquake. *J. Geophys. Res.* **102**, 11765–11774 (1997).

**Acknowledgements** This work was done in the framework of projects ZIP (reference 604713), funded by the E.C. in the call for proposals FP7-PEOPLE-2013-ITN and FRAME (reference CTM2015-71766-R), funded by the Spanish Plan of Research and Innovation. We thank D. Klaeschen and R. von Huene (Geomar) for providing a copy of the P849 seismic data displayed in Fig. 1, T. Lay for his review, and J.-P. Ampuero for his comments on a preliminary version of the work.

**Author contributions** V.S. had the original idea, conceived the physical model, selected and digitized the P-wave velocity profiles, performed the calculations, made the figures except Fig. 1, and wrote the first draft of the manuscript. C.R.R. made the geological interpretation of the physical model, contributed to identifying its implications, processed and pre-stack depth-migrated Java 07 and interpreted both seismic images on Fig. 1, and contributed to writing the manuscript.

**Competing interests** The authors declare no competing interests.

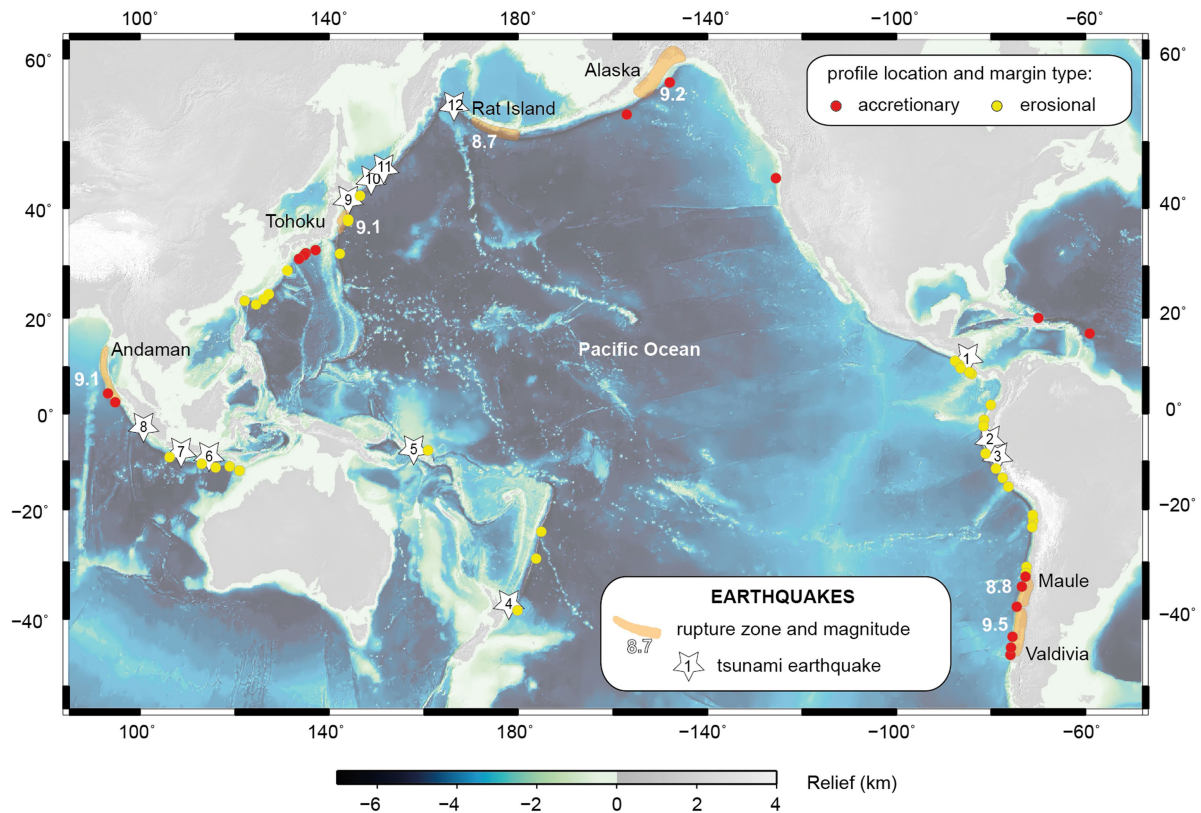
## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1784-0>.

**Correspondence and requests for materials** should be addressed to V.S.

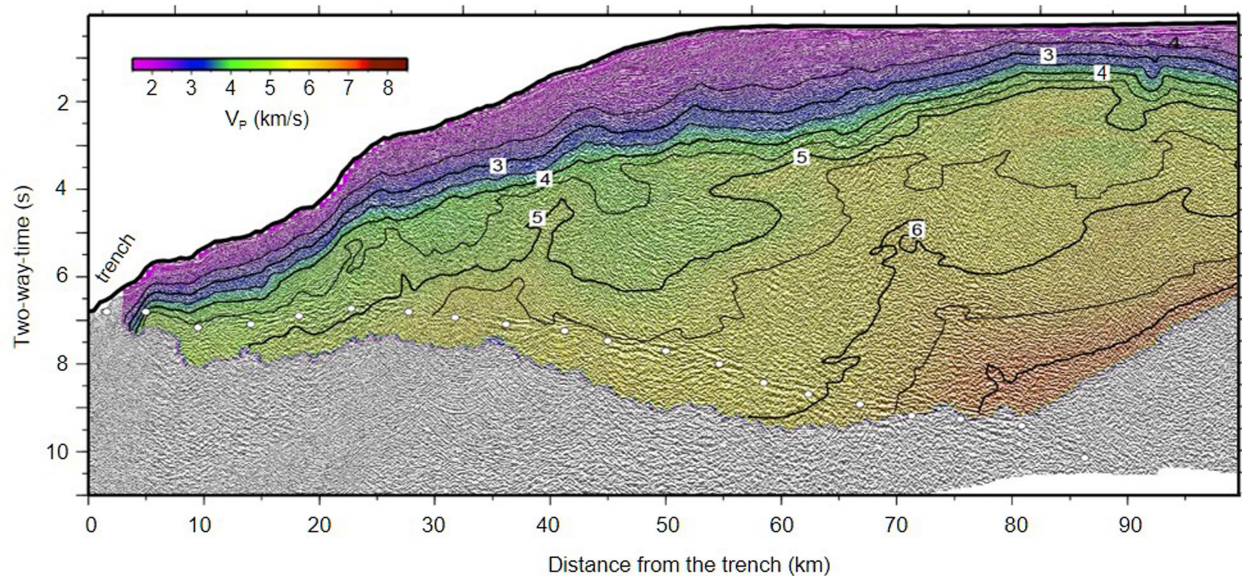
**Peer review information** *Nature* thanks Thorne Lay and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



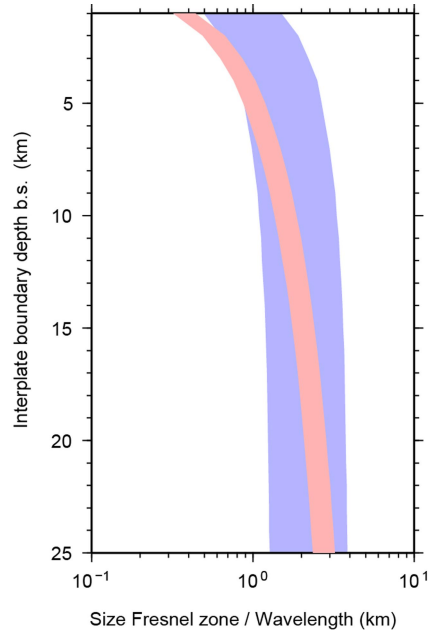
**Extended Data Fig. 1 | Location map of seismic profiles and recent great and tsunami earthquakes.** Colour-coded relief map of seafloor (blue-green) and emerged land (grey). Circles indicate location of the trench-crossing refraction and wide-angle reflection seismic (WAS) profiles used in this study. Yellow-filled circles are in erosional margins and red-filled circles in accretionary margins. The location, type of margin and references for all profiles are listed in Extended Data Table 1. Numbered white stars show locations of 12 events recognized as tsunami earthquakes according to the definition of Kanamori<sup>11</sup>: (1)  $M_w$  7.6, 1992 Nicaragua; (2)  $M_w$  7.6, 1960 Peru; (3)  $M_w$  7.5, 1995 Peru; (4)  $M_s$  7.2, 1947 Hikurangi; (5)  $M_w$  7.1, 2010 Solomon; (6)  $M_w$  7.6, 1994 Java; (7)  $M_w$  7.8, 2006

Java; (8)  $M_w$  7.8, 2010 Mentawai; (9)  $M_w$  8.0, 1896 Sanriku; (10)  $M_w$  7.5, 1975 Kurile; (11)  $M_w$  7.8, 1963 Kurile; (12)  $M_w$  8.2, 1946 Aleutian. Orange polygons display the rupture areas of the six largest megathrust earthquakes since 1960:  $M_w$  9.5, 1960 Valdivia;  $M_w$  9.2, 1964 Alaska;  $M_w$  9.1, 2004 Andaman Islands;  $M_w$  9.1, 2010 Tohoku-Oki;  $M_w$  8.8, 2010 Maule;  $M_w$  8.7, 1965 Rat Island. Hypocentral location, date, magnitude and references for all these earthquakes are listed in Extended Data Table 2. This figure has been created using the GMT software package<sup>46</sup>. The topographic and bathymetric relief data have been taken from the GEBCO Digital Atlas data set<sup>47</sup>. Authors are not aware of any disputed territories shown.



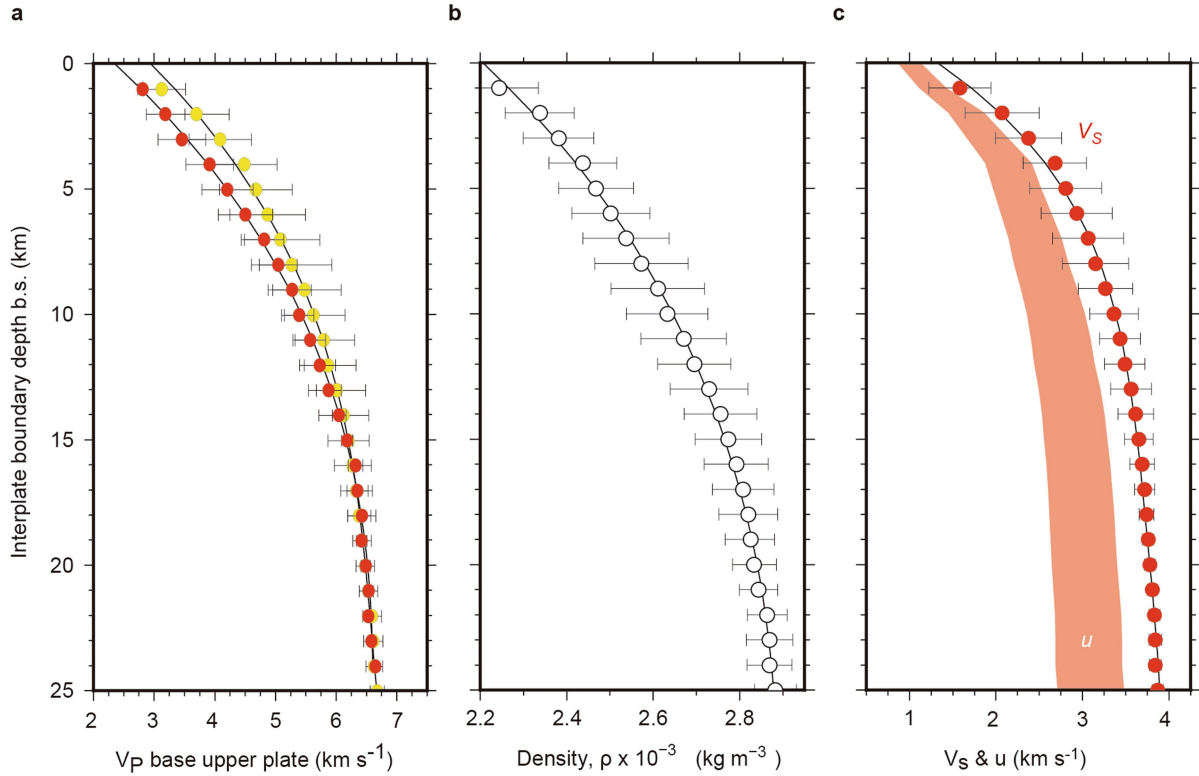
**Extended Data Fig. 2 | Superposition of multichannel seismic image and P-wave velocity model.** Example of superposition of a  $V_p$  model (colour, see scale) on a spatially coincident multichannel seismic image (shading) along profile NIC-20, acquired in the convergent margin of Nicaragua. This profile

crosses the rupture area of the 1992 Nicaragua tsunami earthquake (Extended Data Fig. 1). Black lines show isovelocity contours with their corresponding velocity values. White circles indicate the approximate location of the interplate boundary, where megathrust earthquakes take place.



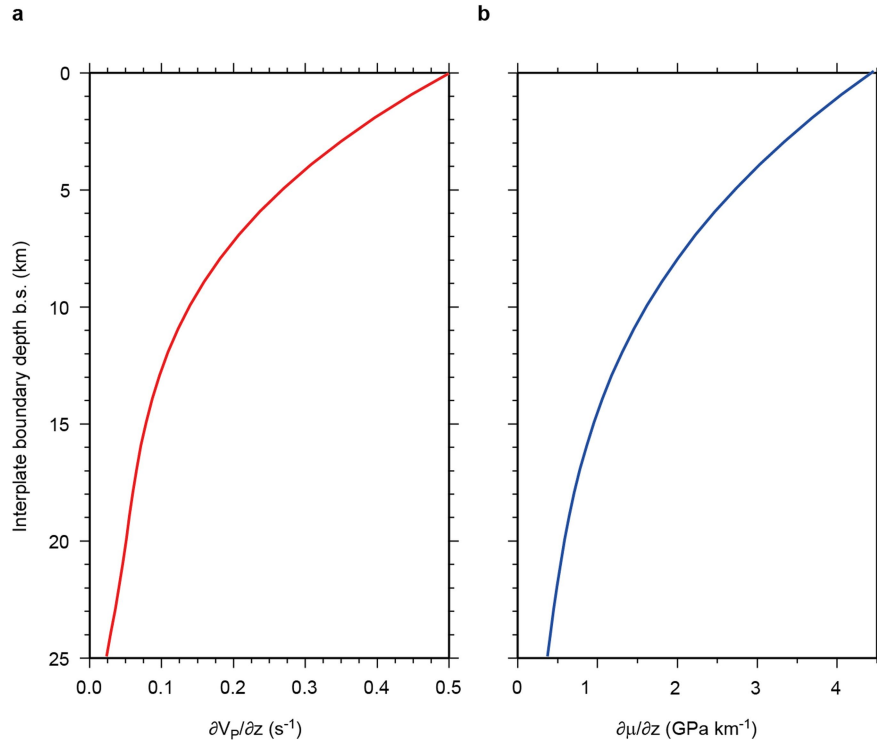
**Extended Data Fig. 3 | Resolution of  $V_p$  models and wavelength of rupture stress wavefield.** The light red polygon displays the width of the Fresnel zone as a function of depth, assuming  $V_p(z)$  in Fig. 2a and energy sources with minimum (maximum) peak frequency  $f_s = 8$  Hz (12 Hz). The blue-lilac polygon indicates the approximate wavelength of the stress wavefield associated to earthquake rupture propagation ( $\lambda_w$ ), assuming  $V_s(z)$  in Extended Data Fig. 5c as propagation velocity, and near-field ground motion spectra with minimum (maximum) peak frequency  $f_{sw} = 1$  Hz (4 Hz) (see Methods for details).





**Extended Data Fig. 4 | Physical properties versus interplate boundary depth.** **a,** Red (yellow) circles show  $V_p$  as a function of  $z$ . It is obtained by averaging digitized  $V_p$  values of accretionary and erosional margins (red and yellow circles, respectively, in Fig. 2b). **b,** White circles show density ( $\rho$ ) just above the interplate boundary, as a function of  $z$ , obtained by applying Brocher's  $\rho(V_p)$  relationship<sup>19</sup>. **c,** Red circles show shear-wave velocity ( $V_s$ ) just

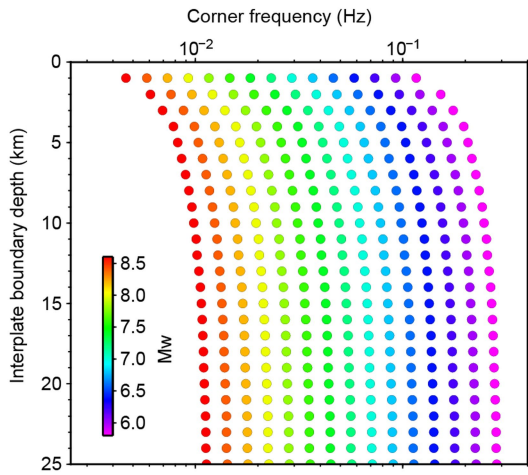
above the interplate boundary, as a function of  $z$ , obtained by applying the  $V_s(V_p)$  relationship from Brocher (ref. <sup>19</sup>). The shaded polygon covers the range of possible mode III rupture velocities, as a function of  $z$ , according to field observations:  $u(z) = (0.7-0.9)V_s(z)$ . The black line is a fourth-order polynomial regression fit of the  $V_p(z)$ ,  $\rho(z)$  and  $V_s(z)$  values, respectively. The size of the error bars in all cases is one standard deviation.



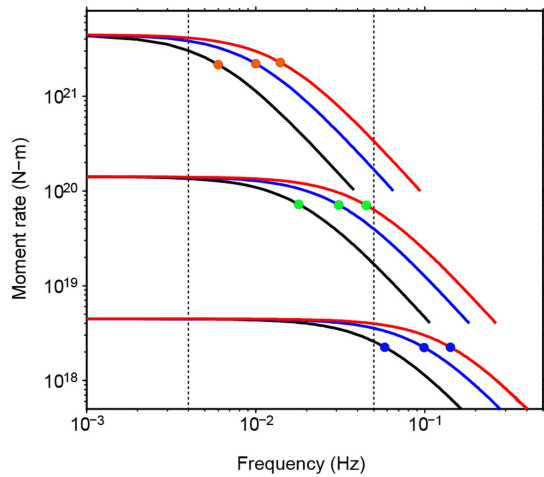
**Extended Data Fig. 5 | P-wave velocity and rigidity gradients.** **a.** Red line shows the depth gradient of  $V_p$  as a function of interplate boundary depth,  $\partial V_p(z)/\partial z$ . It corresponds to the derivative of the  $V_p(z)$  polynomial regression fit

(black line in Fig. 2c). **b.** Blue line shows the depth gradient of  $\mu$  as a function of interplate boundary depth,  $\partial \mu(z)/\partial z$ . It corresponds to the derivative of the  $\mu(z)$  polynomial regression fit (black line in Fig. 2d).

a

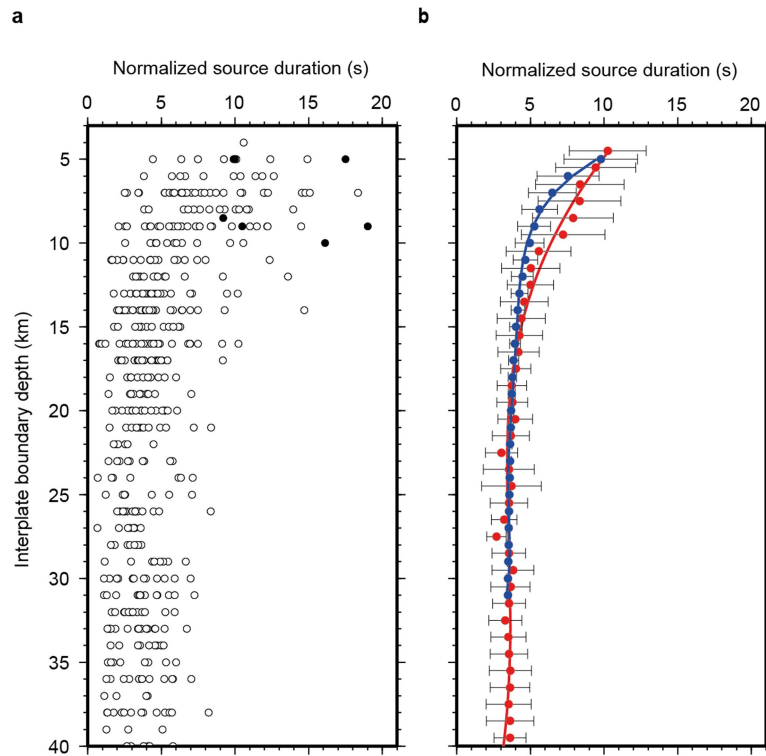


b



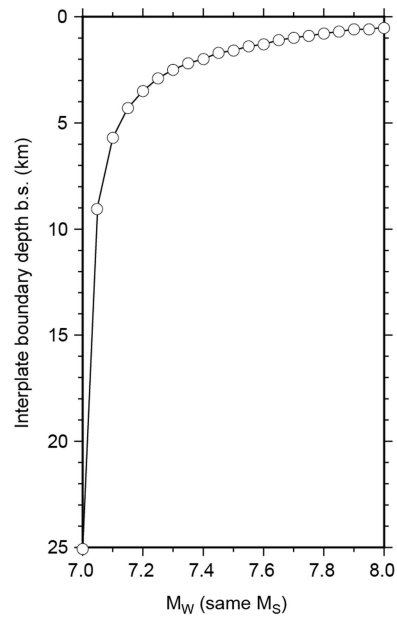
### Extended Data Fig. 6 | Corner frequency and moment rate spectra.

**a**, Coloured circles show corner frequency as a function of interplate boundary depth ( $z$ ), for events of  $M_w = 5.8$ – $8.6$ .  $\Delta\sigma = 3$  MPa is used in the calculations, and  $V_s(z)$  is taken from Extended Data Fig. 3c. The colour scale indicates  $M_w$ . **b**, Solid lines show calculated moment rate spectra for three events of  $M_w = 6.4$  (bottom),  $7.4$  (mid) and  $8.4$  (top). Black, blue and red lines correspond to  $V_s$  at  $z = 1$  km,  $6$  km and  $25$  km, respectively, for each event. Coloured circles indicate corner frequency according to colour code in Extended Data Fig. 6a. Vertical dashed lines indicate periods of  $250$  s and  $20$  s, reference to calculate  $M_w$  and  $M_s$ , respectively. In all cases, note the high-frequency depletion at shallow depths.



**Extended Data Fig. 7 | Source duration of circum-Pacific megathrust earthquakes. a,** White circles show scaled source duration of 525 moderate size ( $M_w$  5.0–7.5) shallow megathrust subduction earthquakes from around the circum-Pacific, as a function of depth. Black circles show the same source parameters for six large tsunami earthquakes. Data are from ref. <sup>21</sup>. This article contains all the information on the procedure followed to calculate the source

parameters. **b,** Red circles are the normalized source duration in **a** averaged within a 2-km-thick sliding window. Blue circles correspond to relative rupture duration in Fig. 3b scaled to fit average normalized rupture duration within the regular domain in **a** (approximately 3.5 s), and shifted 4.5 km down to compensate for the difference between depth below sea surface and depth below seafloor. Error bars are one standard deviation.



**Extended Data Fig. 8 | Range of variation of  $M_w$  for a given  $M_s$ .** White circles show  $M_w$  for events occurring at different interplate boundary depth b.s. that have the same spectral amplitude at 20 s (hence equivalent  $M_s$ ).



**Extended Data Table 1 | Location of seismic profiles and margin type**

Number	Longitude	Latitude	Type	Region	Reference
1	-86.7	10.3	E	Nicaragua	41
2	-84.13	8.47	E	Costa Rica	48
3	-80.0	1.85	E	Ecuador	49
4	-71.1	-22.2	E	North Chile	50
5	-71.3	-23.4	E	North Chile	50
6	-73.5	-34.5	A	Central Chile	51
7	-75.85	-45.55	A	South Chile	52
8	-175.1	-24.4	E	Tonga	53
9	94.7	2.45	A	Sumatra	54
10	124.5	22.6	E	South Ryukyu	55
11	135.1	32.3	A	Nankai	56
12	144.0	37.9	E	Tohoku	57
13	-125.5	44.7	A	Cascadia	58
14	-74.6	-38.0	A	South Chile	38
15	-81.1	-8.5	E	North Peru	59
16	-79.0	-11.6	E	Central Peru	59
17	-77.5	-13.5	E	Central Peru	59
18	-76.5	-15.3	E	South Peru	60
19	113.0	-10.65	E	Java	61
20	179.9	-38.6	E	New Zealand	62
21	160.9	-7.8	E	Solomon	63
22	142.25	32.25	E	Izu Bonin	64
23	-75.75	-44.5	A	South Chile	52
24	-87.6	11.2	E	Nicaragua	65
25	-86.4	9.7	E	Costa Rica	66
26	183.8	-29.5	E	Kermadec	67
27	122.15	23.35	E	Taiwan	68
28	106.3	-9.25	E	Java	69
29	116.0	-11.4	E	Lombok	70
30	118.9	-11.2	E	Sumba	70
31	-148.0	57.4	A	Alaska	71
32	-72.5	-31.0	E	Central Chile	39
33	146.6	42.0	E	Kuril	72
34	137.1	32.9	A	Tonankai	73
35	134.75	31.9	A	Shikoku	74
36	93.2	4.3	A	Sumatra	75
37	-72.65	-32.1	E	Central Chile	39
38	-59.1	16.8	A	Antilles	76
39	-84.5	8.75	E	Costa Rica	77
40	-71.2	-21.1	E	North Chile	78
41	-81.5	-1.3	E	Ecuador	79
42	-81.6	-2.6	E	Ecuador	80
43	-75.5	-42.8	A	South Chile	52
44	-157.0	53.5	A	Alaska	81
45	144.0	38.1	E	Miyagi	82
46	131.5	29.25	A	North Rykyu	83
47	127.5	24.7	E	Central Rykyu	83
48	126.6	23.95	E	South Rykyu	83

Geographical locations of the 48 wide-angle seismic profiles along the circum-Pacific included in the data set. The type of margin is indicated in the fourth column by an A (accretionary) or an E (erosional), which correspond to the red and yellow circles in Extended Data Fig. 1, respectively. Data taken from refs. <sup>39,41-83</sup>, Reference number is indicated in the last column.

Extended Data Table 2 | Location and magnitude of circum-Pacific megathrust earthquakes

Number	Longitude	Latitude	Day	Month	Year	M <sub>s</sub>	M <sub>w</sub>	Region	Reference
Largest Megathrust Earthquakes									
1	-73.407	-38.143	22	5	1960	-	9.5	Valdivia	84
2	-147.34	60.91	28	3	1964	-	9.2	Alaska	85
3	95.982	3.295	26	12	2004	-	9.1	Andaman	86
4	142.373	38.297	11	3	2011	-	9.1	Tohoku	87
5	-72.898	-36.122	27	2	2010	-	8.8	Maule	98
6	178.715	51.251	4	2	1965	-	8.7	Rat Island	89
Tsunami Earthquakes									
1	-87.38	11.75	2	9	1992	7.0	7.6	Nicaragua	12
2	-80.90	-6.72	20	11	1960	6.75	7.6	Peru	22
3	-80.23	-9.95	21	2	1996	-	7.5	Peru	90
4	178.9	-38.8	25	3	1947	7.2	7.1	Hikurangi	91
5	157.3	-8.3	3	1	2010	-	7.1	Solomon	92
6	113.04	-10.28	2	6	1994	7.2	7.6	Java	27
7	107.39	-9.26	17	7	2006	7.2	7.8	Java	93
8	100.14	-3.49	25	10	2010	7.1	7.8	Mentawai	23
9	144.0	39.5	15	6	1896	7.2	8.0	Sanriku	94
10	147.63	43.07	10	6	1975	7.0	7.5	Kurile	22
11	150.7	44.7	20	10	1963	7.2	7.8	Kurile	22
12	163.1	53.32	1	4	1946	7.4	8.2	Aleutian	95

Geographical location, date, and magnitudes of the circum-Pacific megathrust earthquakes shown in Extended Data Fig. 1. The list includes the six largest-magnitude earthquakes that have occurred since 1960, as well as 12 events identified as tsunami earthquakes. Data taken from refs. <sup>12,22,23,84–95</sup>. Reference numbers for the source parameters and energy release characteristics of all the events are indicated in the last column.

# Cretaceous fossil reveals a new pattern in mammalian middle ear evolution

<https://doi.org/10.1038/s41586-019-1792-0>

Haibing Wang<sup>1,2</sup>, Jin Meng<sup>3</sup> & Yuanqing Wang<sup>1,2,4\*</sup>

Received: 23 April 2019

Accepted: 23 October 2019

Published online: 27 November 2019

The evolution of the mammalian middle ear is thought to provide an example of ‘recapitulation’—the theory that the present embryological development of a species reflects its evolutionary history. Accumulating data from both developmental biology and palaeontology have suggested that the transformation of post-dentary jaw elements into cranial ear bones occurred several times in mammals<sup>1,2</sup>. In addition, well-preserved fossils have revealed transitional stages in the evolution of the mammalian middle ear<sup>1,3,4</sup>. But questions remain concerning middle-ear evolution, such as how and why the post-dentary unit became completely detached from the dentary bone in different clades of mammaliaforms. Here we report a definitive mammalian middle ear preserved in an eobaatarid multituberculate mammal, with complete post-dentary elements that are well-preserved and detached from the dentary bones. The specimen reveals the transformation of the surangular jaw bone from an independent element into part of the malleus of the middle ear, and the presence of a restricted contact between the columelliform stapes and the flat incus. We propose that the malleus–incus joint is dichotomic in mammaliaforms, with the two bones connecting in either an abutting or an interlocking arrangement, reflecting the evolutionary divergence of the dentary–squamosal joint<sup>4</sup>. In our phylogenetic analysis, acquisition of the definitive mammalian middle ear in allotherians such as this specimen was independent of that in monotremes and therians. Our findings suggest that the co-evolution of the primary and secondary jaw joints in allotherians was an evolutionary adaptation allowing feeding with unique palinal (longitudinal and backwards) chewing. Thus, the evolution of the allotherian auditory apparatus was probably triggered by the functional requirements of the feeding apparatus.

Mammalia Linnaeus, 1758

Multituberculata Cope, 1884

Eobaataridae Kielan-Jaworowska, Dashzeveg and Trofimov, 1987

*Jeholbaatar kielanae* gen. et sp. nov.

**Etymology.** *Jehol* derives from the Jehol Biota ecosystem of Cretaceous northeastern China; *baatar* (Mongolian), meaning hero, is a common suffix for Asian Cretaceous multituberculate names; *kielanae* honours the Polish palaeontologist Zofia Kielan-Jaworowska for her contribution to the study of multituberculates.

**Holotype.** A nearly complete skeleton (IVPP V20778; Fig. 1), housed in the Institute of Vertebrate Paleontology and Paleoanthropology, Beijing, China.

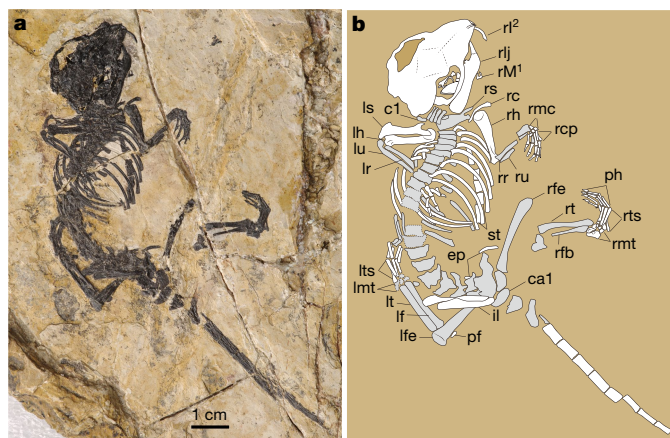
**Locality and age.** The specimen is from the Jiufotang Formation near Changzigou, Lingyuan City, Liaoning Province, China, dated to approximately 120 million years ago<sup>5</sup>.

**Diagnosis.** Dental formula of  $I^3 \cdot C^0 \cdot P^5 \cdot M^2 / I_1 \cdot C_0 \cdot P_3 \cdot M_2$  (I, incisor; C, canine; P, premolar; M, molar; superscript and subscript denote upper and

lower teeth, respectively), with the following multituberculate characteristics (Extended Data Figs. 1, 2): cranium dorsoventrally compressed; masseteric fossa anteriorly extending below lower premolars; lingual offset of  $M^2$  relative to  $M^1$ ; enlarged single lower incisor; blade-like  $P_4$ ; definitive mammalian middle ear (Extended Data Figs. 3, 4). Among multituberculates, *Jeholbaatar* is referable to eobaatarids on the basis of: upper canines absent;  $I^3$  transversely wide; and eight serrations and a posterolabial cusp on  $P_4$ . *Jeholbaatar* differs from most eobaatarids (except for *Eobaatar* and *Heishanbaatar*) in having eight serrations on  $P_4$ ; differs from *Eobaatar* in having reduced  $P_{2-3}$ , more buccal cusps on  $M^1$ , and a ridged cusp row on  $P^5$ ; differs from *Heishanbaatar* in having an oval lateral outline of  $P_3$  and more cusps of lower molars; and differs from *Sinobaatar* in having a posterior cuspule on  $I^2$ , two cusp rows of  $P^5$ , and different cusp counts of upper and lower molars.

Phylogenetic analyses place *Jeholbaatar* within the monophyletic eobaatarids and closely related to *Sinobaatar* (Extended Data Figs. 5, 6). The body mass of *Jeholbaatar* is estimated to be approximately 50 g on the basis of its skull length<sup>6</sup> (see Supplementary Information).

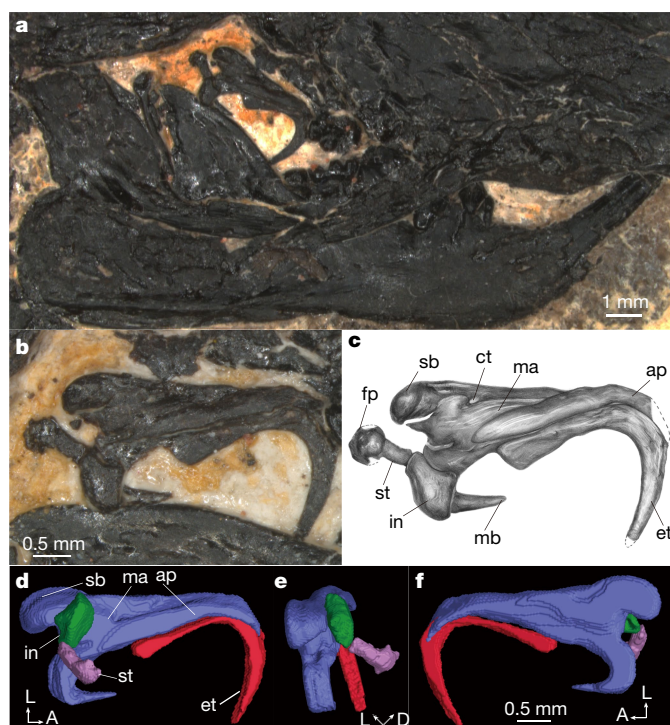
<sup>1</sup>Key Laboratory of Vertebrate Evolution and Human Origins of Chinese Academy of Sciences, Institute of Vertebrate Paleontology and Paleoanthropology, Beijing, China. <sup>2</sup>Centre for Excellence in Life and Palaeoenvironment, Chinese Academy of Sciences, Beijing, China. <sup>3</sup>Division of Paleontology, American Museum of Natural History, New York, NY, USA. <sup>4</sup>College of Earth and Planetary Sciences, University of Chinese Academy of Sciences, Beijing, China. \*e-mail: wangyuanqing@ivpp.ac.cn



**Fig. 1** | The Cretaceous multituberculate *Jeholbaatar kielanae*. **a**, Holotype (IVPP V20778) in dorsal view. **b**, Line drawing of the holotype. Grey shading indicates damaged elements. ca1, first caudal vertebra; c1, atlas; ep, epipubis; il, ilium; lf, left fibula; lfe, left femur; lh, left humerus; lmt, left metatarsals; lr, left radius; ls, left scapulocoracoid; lt, left tibia; lts, left tarsals; lu, left ulnar; pf, parafibula; ph, phalanges; rc, right clavicle; rcp, right carpals; rh, right humerus; rfb, right fibula; rfe, right femur; rl<sup>2</sup>, right I<sup>2</sup>; rl, right lower jaw; rmc, right metacarpals; rmt, right metatarsals; rM<sup>1</sup>, right M<sup>1</sup>; rr, right radius; rs, right scapulocoracoid; rt, right tibia; ru, right ulna; st, sternum.

*Jeholbaatar* is inferred to be scansorial on the basis of its manual and pedal morphology, and the phalangeal index of its third digits is the greatest among multituberculates, with postcranials preserved (Fig. 1, Extended Data Fig. 7 and Extended Data Table 1). Given the morphology of the lower cheek teeth, we infer that *Jeholbaatar* is similar to *Eobaatar* in having an omnivorous diet, feeding on arthropods, worms and plant items<sup>7</sup>. The palinal jaw joint<sup>4</sup> and the distinct attachment for the masseter muscle suggest a unique palinal jaw movement while chewing (Extended Data Fig. 1).

The well-preserved left middle-ear bones are mediadorsally exposed and articulated nearly in anatomical position (Fig. 2 and Extended Data Figs. 3, 4). This unit is clearly detached from the dentary, as indicated by the absence of a sulcus on the lingual side of the dentary, as in other multituberculates; thus, *Jeholbaatar* has by definition the definitive mammalian middle ear (DMME)<sup>8</sup>. The stapes is columelliform—microperforate and distinct from the typical columelliform stapes of *Lambdopsalis*<sup>9</sup> in having a robust shaft, a less expanded stapelial footplate (the proximal end of the stapes), and a more basally positioned stapelial foramen (Extended Data Fig. 3). Laterally, the stapelial head—not exposed fully in dorsal view—is narrowed (relative to the proximal end) as in therian mammals, and articulates with the stapelial process (the long crus) of the incus through a restricted contact (relative to the broad end-on contact in other mammaliaforms<sup>10</sup>), preserving no sign of the extrastapes. The complete incus, previously unknown in a multituberculate, is slightly displaced ventrally, revealing its shape and orientation. The incus body is flat and lies medial to the transverse portion of the malleus body. Its morphology and small size suggest that the incus may not contact the squamosal dorsally. The proximal portion of the anterior process of the malleus is dorsally thicker than the transverse portion of the malleus body. A foramen, presumably for the chorda tympani, perforates the malleus (Extended Data Fig. 3). The malleus body bears a short manubrium projecting anteroventrally. The ventrolateral part of the malleus is thick and wedge-shaped, which we interpret as the remnant of the surangular. It consists of an anterior projection and a convex posterior end (surangular boss; Fig. 2). The posterior portion of the surangular extends posterolaterally and the posterodorsal surface of the surangular boss remains smooth and restricted medially by a distinct neck, which is reminiscent of an articular surface. The ectotympanic is large and roughly sickle-shaped



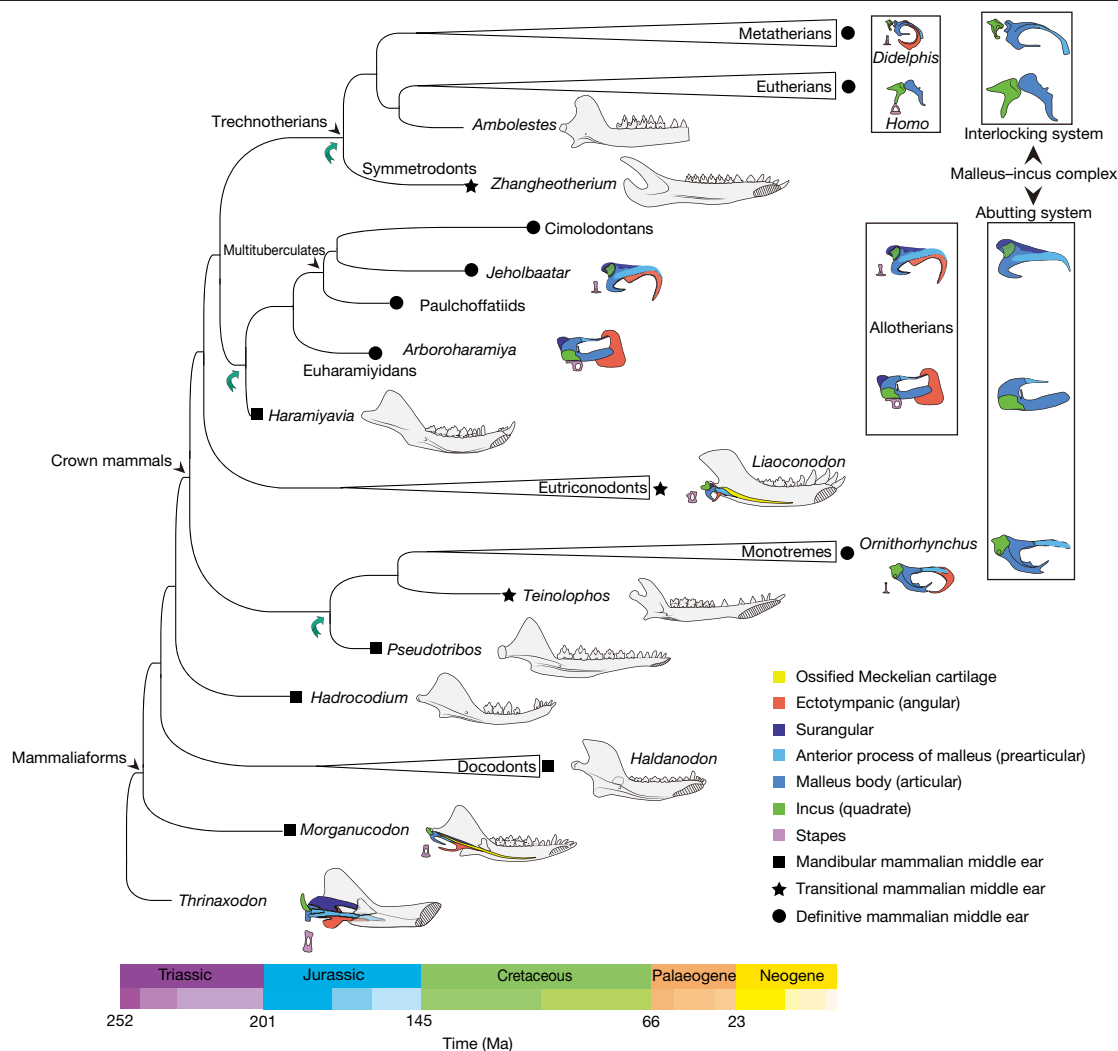
**Fig. 2** | Middle-ear bones of *Jeholbaatar kielanae*. **a**, Left middle-ear bones, slightly displaced between the right lower jaw and braincase. **b**, Left middle ear bones exposed in dorsal view. **c**, Interpretative drawing of left ear bones. **d–f**, Reconstructions of left middle ear bones, showing articulation of these elements, in dorsal (**d**), posterior (**e**) and ventral (**f**) views; the surangular, the malleus body and the anterior process of the malleus are combined as a unit. A, anterior; ap, anterior process of malleus; ct, foramen for chorda tympani; D, dorsal; et, ectotympanic; fp, footplate of stapes; in, incus; L, lateral; ma, body of malleus; mb, manubrium of malleus; sb, surangular boss; st, stapes.

with a gently posteriorly curved ventral limb and no anterior limb. The posterior portion of the horizontal limb is slightly expanded medially. The ectotympanic connects firmly to the malleus, suggesting that they may function as a unit.

The specimen provides important evidence regarding mammalian middle-ear evolution, revealing a unique configuration with more complete and complex components than those reported previously in Cretaceous multituberculates<sup>11</sup>. Under our phylogenetic framework, the DMME has evolved independently at least three times, in allotherians, monotremes and trechnotherians (Fig. 3).

Detachment of the auditory bones from the dentary was accompanied by loss of the anterior limb of the ectotympanic during development of the DMME, which evolved in parallel in monotremes, therians and allotherians (*Arboroharamiya* and *Jeholbaatar*)<sup>4,12</sup>. The hook-like ectotympanic is plesiomorphic for early mammals, as demonstrated by *Arboroharamiya* and *Jeholbaatar*, contrasting with the ring-like form of the Early Cretaceous *Amolestes*<sup>13</sup>.

The incus–stapes complex has been simplified in *Jeholbaatar* through a reduction in size and restricted incus–stapes contact. Whether the rod-like or the asymmetric bicurrate form represents the ancestral morphotype of the mammaliaform stapes is still disputed<sup>14</sup>. *Jeholbaatar* reveals a transitional stage in the evolution of the stapes, intermediate between the rod-like form (observed in cynodonts, *Arboroharamiya* and *Chaoyangodens*<sup>4,10</sup>) and the typical columelliform morphology (with a slender shaft, as seen in *Lambdopsalis*<sup>9</sup>). Although there are different interpretations of some previously reported multituberculate stapes<sup>14</sup>, the robust shaft and less expanded footplate of the stapes in *Jeholbaatar* is distinct from the asymmetric bicurrate morphology (observed in *Pseudobolodon*<sup>14</sup>). This suggests several processes for



**Fig. 3 | Evolution of the mammalian middle ear in different mammaliaform clades.** This simplified phylogeny is based on the strict consensus of parsimony analysis (see Extended Data Fig. 5 and Supplementary Information). The green arrows denote independent evolution of the DMME in mammaliaforms. In the second column from the right are the middle-ear bone complexes of different taxa; at the right are the corresponding diagnostic

configurations of the abutting and interlocking systems of the malleus-incus complex: in the abutting system, the malleus and incus contact dorsoventrally; in the interlocking system, this contact is rostrocaudal. Reconstructions of left middle ears are taken from the literature (see Methods and Supplementary Information). Graded colours in the key at the bottom denote Early, Middle and Late periods. Ma, million years ago.

evolution of the stapes in mammaliaforms, with independent acquisition of a bicurate morphotype in *Pseudobolodon* and *Kryptobaatar*. The restricted incus-stapes contact of *Jeholbaatar* is derived by comparison with other mammaliaforms that have a broad end-on contact between these two bones. The development of the stapedia process of the incus, as an out-lever of the lever system during sound transition, is beneficial for the amplification of airborne sound<sup>1</sup>.

Identification of the surangular in *Jeholbaatar* reinforces the argument that the remnant of the ancient 'reptilian' element exists in crown mammals (allotherians)<sup>4</sup>. It also fills a gap in the fossil record of the transformation of the surangular from an independent element to an accessory of the malleus<sup>3,4</sup>, providing clues to the evolution of the surangular in mammaliaforms. In *Jeholbaatar*, the manubrium of the malleus is short and gradually tapers anterolaterally from the malleus body. This is the plesiomorphic condition, lacking the clear distinction between the manubrial base and the manubrium observed in other known Mesozoic mammaliaforms that have preserved the middle-ear bones (except *Arboroharamiya*)<sup>3</sup>. *Jeholbaatar* also provides evidence of a thickened malleus in a Mesozoic mammaliaform. This condition is defined specifically by the expression of the *Bapx1* gene in mice<sup>15</sup>, implying similar embryonic development in *Jeholbaatar*.

The malleus-incus complex of *Jeholbaatar* is similar to that of *Arboroharamiya* and extant monotremes<sup>4,12,16</sup>, with a dorsoventral contact of the malleus-incus complex. Given that the mammaliaform malleus-incus complex is derived from the primary joint in lower tetrapods, this raises an interesting issue concerning how the incus shifted dorsal to the malleus body during the transformation of the middle-ear bones. We propose that the articular configuration of the malleus-incus complex is dichotomic among mammaliaforms: the abutting system is characterized by a dorsoventral contact, as observed in monotremes, *Arboroharamiya*, and *Jeholbaatar*<sup>4,16</sup>; and the interlocking system has a rostrocaudal contact (and later hinge-like articulation), as observed in *Morganucodon*<sup>17</sup>, *Liaconodon* and other mammals except allotherians and monotremes<sup>3</sup>. This interpretation of the malleus-incus articulation contradicts previous proposals regarding other multituberculates<sup>14,18</sup>. However, in light of the unequivocal articulated middle-ear bones in *Jeholbaatar*, we postulate that the abutting system persisted in later multituberculate evolution. Whether this configuration is consistent in allotherians that have a mandibular mammalian middle ear (such as *Haramiyavia*) and transitional mammalian middle ear remains unknown. It has been proposed that the primary joint (malleus and incus) in mammals is determined by members of the *Gdf* gene family



(*Gdf5* and *Gdf6*)<sup>15</sup>. If all of these hypotheses are correct, then the developmental divergence of the primary joint (as reflected in the malleus–incus articulation) in mammaliaforms occurred deep in the Middle to Late Jurassic period, resulting in a shift in the position of the incus dorsal over the malleus (Extended Data Fig. 4). Despite the morphological distinction of the middle-ear bones between *Jeholbaatar* and *Arboroharamiya*, the configuration of the abutting system is coincident with the palinal jaw joint in multituberculates and euharamiyids. The timing of the divergence of malleus–incus configurations (the abutting and interlocking systems) and the dichotomic morphotype of the squamosal–dentary jaw joint (palinal and hinge-like)<sup>4</sup> supports the hypothesis that the primary and secondary jaw joints co-evolved in allotherians.

The evolution of the DMME is associated with morphogenetic processes in the post-dentary bones, and causes of the detachment of Meckel’s cartilage are hierarchical<sup>1,19–21</sup>. Palaeontological and developmental findings have rendered two conventional hypotheses for the degeneration of Meckel’s cartilage (the brain-expansion hypothesis<sup>22</sup> and negative ontogenetic allometry of the middle-ear bones<sup>23</sup>) less plausible<sup>20,24–26</sup>. Instead, given the evidence from *Arboroharamiya* and *Jeholbaatar*, the evolution of the DMME in allotherians might be explained by biomechanical functional constraints during feeding<sup>27,28</sup>, with co-evolution of the primary and secondary jaw joints being an adaptation for the unique palinal chewing of allotherians. Earlier acquisition of the DMME in allotherians also implies a shortened transitional mammalian middle-ear stage. The abutting-system configuration permitted longitudinal and vertical reduction of the middle-ear bones in some mammaliaforms. Detachment of the middle-ear bones (followed by better handling of biomechanical loads related to mastication on the medial side of the dentary<sup>29</sup>) and the abutting-system configuration could have increased the degree of food comminution per palinal power stroke in those allotherians with the DMME, and reduced the impact of feeding on the hearing apparatus. As such, selective pressure to detach the middle-ear bones (the hearing apparatus) in order to increase feeding efficiency could have been stronger in allotherians than in clades characterized by the interlocking system, showing that feeding was an important trigger in DMME evolution.

The homoplastic evolution of the DMME observed in fossils is consistent with developmental evidence, revealing diverse mechanisms for the detachment of Meckel’s cartilage in different lineages<sup>20</sup>. The presence of the surangular remnant in *Jeholbaatar* might represent a recapitulation of the ancestral state, and suggests that evolution of the DMME could be an instance of von Baer’s law of embryology<sup>30</sup>—although this hypothesis requires further investigation in a developmental context.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1792-0>.

1. Luo, Z. X. Developmental patterns in Mesozoic evolution of mammal ears. *Annu. Rev. Ecol. Syst.* **42**, 355–380 (2011).

2. Maier, W. & Ruf, I. Evolution of the mammalian middle ear: a historical review. *J. Anat.* **228**, 270–283 (2016).
3. Meng, J., Wang, Y. & Li, C. Transitional mammalian middle ear from a new Cretaceous Jehol eutriconodont. *Nature* **472**, 181–185 (2011).
4. Han, G., Mao, F., Bi, S., Wang, Y. & Meng, J. A Jurassic gliding euharamiyid mammal with an ear of five auditory bones. *Nature* **551**, 451–456 (2017).
5. He, H. Y. et al. Timing of the Jiufotang Formation (Jehol Group) in Liaoning, northeastern China, and its implications. *Geophys. Res. Lett.* **31**, L12605 (2004).
6. Millien, V. & Bovy, H. When teeth and bones disagree: body mass estimation of a giant extinct rodent. *J. Mamm.* **91**, 11–18 (2010).
7. Wilson, G. P. et al. Adaptive radiation of multituberculate mammals before the extinction of dinosaurs. *Nature* **483**, 457–460 (2012).
8. Allin, E. F. & Hopson, J. in *The Evolutionary Biology of Hearing* Ch. 28 (eds Webster, D. B. et al.) 587–614 (Springer, 1992).
9. Meng, J. The stapes of *Lambdopsalis bulla* (Multituberculata) and transformational analyses on some stapedial features in Mammaliaformes. *J. Vertebr. Paleontol.* **12**, 459–471 (1992).
10. Meng, J. & Hou, S. L. Earliest known mammalian stapes from an Early Cretaceous eutriconodontan mammal and implications for the evolution of mammalian middle ear. *Palaeontol. Pol.* **67**, 181–196 (2016).
11. Rougier, G. W., Wible, J. R. & Novacek, M. J. Middle-ear ossicles of the multituberculate *Kryptobaatar* from the Mongolian Late Cretaceous: implications for mammalian relationships and the evolution of the auditory apparatus. *Am. Mus. Novit.* **3187**, 1–43 (1996).
12. Fleischer, G. Evolutionary principles of the mammalian middle ear. *Adv. Anat. Embryol. Cell Biol.* **55**, 3–70 (1978).
13. Bi, S. et al. An Early Cretaceous eutherian and the placental–marsupial dichotomy. *Nature* **558**, 390–395 (2018).
14. Schultz, J. A., Ruf, I. & Martin, T. Oldest known multituberculate stapes suggests an asymmetric biclural pattern as ancestral for Multituberculata. *Proc. R. Soc. B* **285**, 20172779 (2018).
15. Tucker, A. S., Watson, R. P., Lettice, L. A., Yamada, G. & Hill, R. E. Bapx1 regulates patterning in the middle ear: altered regulatory role in the transition from the proximal jaw during vertebrate evolution. *Development* **131**, 1235–1245 (2004).
16. Zeller, U. in *Mammal Phylogeny: Mesozoic Differentiation, Multituberculates, Monotremes, Early Therians, and Marsupials* Ch. 8 (eds Szalay, F. S. et al.) 95–107 (Springer, 1993).
17. Kermack, K. A., Mussett, F. & Rigney, H. W. The skull of *Morganucodon*. *Zool. J. Linn. Soc.* **71**, 1–158 (1981).
18. Luo, Z. X. et al. New evidence for mammaliaform ear evolution and feeding adaptation in a Jurassic ecosystem. *Nature* **548**, 326–329 (2017).
19. Anthwal, N., Urban, D. J., Luo, Z.-X., Sears, K. E. & Tucker, A. S. Meckel’s cartilage breakdown offers clues to mammalian middle ear evolution. *Nature Ecol. Evol.* **1**, 0093 (2017).
20. Urban, D. J. et al. A new developmental mechanism for the separation of the mammalian middle ear ossicles from the jaw. *Proc. R. Soc. B* **284**, 20162416 (2017).
21. Lautenschlager, S., Gill, P. G., Luo, Z.-X., Fagan, M. J. & Rayfield, E. J. The role of miniaturization in the evolution of the mammalian jaw and middle ear. *Nature* **561**, 533–537 (2018).
22. Rowe, T. Coevolution of the mammalian middle ear and neocortex. *Science* **273**, 651–654 (1996).
23. Zeller, U. in *Morphogenesis of the Mammalian Skull* (eds Kuhn, H.-J. & Zeller, U.) 17–50 (Paul Parey, 1987).
24. Meng, J., Hu, Y. M., Wang, Y. Q. & Li, C. K. The ossified Meckel’s cartilage and internal groove in Mesozoic mammaliaforms: implications to origin of the definitive mammalian middle ear. *Zool. J. Linn. Soc.* **138**, 431–448 (2003).
25. Wang, Y., Hu, Y., Meng, J. & Li, C. An ossified Meckel’s cartilage in two Cretaceous mammals and origin of the mammalian middle ear. *Science* **294**, 357–361 (2001).
26. Ramirez-Chaves, H. E. et al. Mammalian development does not recapitulate suspected key transformations in the evolutionary detachment of the mammalian middle ear. *Proc. R. Soc. Lond. B* **283**, 20152606 (2016).
27. Crompton, A. W. & Parker, P. Evolution of the mammalian masticatory apparatus. *Am. Sci.* **66**, 192–201 (1978).
28. Allin, E. F. Evolution of the mammalian middle ear. *J. Morphol.* **147**, 403–437 (1975).
29. Crompton, A. W. & Hylander, W. L. in *The Ecology and Biology of Mammal-like Reptiles* (eds Hotton, N. et al.) 78–98 (Smithsonian Institution Press, 1986).
30. Abzhanov, A. von Baer’s law for the ages: lost and found principles of developmental evolution. *Trends Genet.* **29**, 712–722 (2013).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

## Methods

### Specimen preparation

At the early stage of preparation, the specimen was mainly exposed in dorsal view. After it was scanned using computed laminography, the skull was prepared from the backside of the slab to expose the skeletal morphology in ventral view.

### Measurements

Skeletal elements were measured in ImageJ.

### Figures

Middle ear reconstructions are based on the following references: *Thrinaxodon*, *Morganucodon* and *Didelphis*, ref. <sup>28</sup>; *Hadrocodium*, refs. <sup>18,31</sup>; *Pseudotribos*, ref. <sup>32</sup>; *Ornithorhynchus*, ref. <sup>16</sup>; *Teinolophos*, ref. <sup>33</sup>; *Liaconodon*, ref. <sup>3</sup>; *Haramiyavia*, ref. <sup>34</sup>; *Arboroharamiya*, ref. <sup>4</sup>; *Zhangheotherium*, ref. <sup>35</sup>; *Amolestes*, ref. <sup>13</sup>; *Haldanodon*, ref. <sup>36</sup>.

### Computed laminography

We carried out scanning using a microcomputed laminography system (developed by the Institute of High Energy Physics, Chinese Academy of Sciences (CAS) at the Key Laboratory of Vertebrate Evolution and Human Origins, CAS). The specimen was scanned with a beam energy of 60 kV and a flux of 40  $\mu$ A at a resolution of 8.7  $\mu$ m per pixel, using a 360° rotation with a step size of 1°. We reconstructed a total of 360 image slices with a size of 2,048  $\times$  2,048 pixels using a modified Feldkamp algorithm developed by the Institute of High Energy Physics, CAS. Three-dimensional reconstruction of the auditory bones and teeth was conducted in VGStudio 3.0.

### Taxonomic terminology

We use the node-based concept for crown clades of Mammalia; the term ‘mammaliaforms’ refers to taxa in Mammaliaformes<sup>37</sup>. Given recent studies<sup>4,38</sup>, we regard Allotheria as a monophyletic group, and we test this hypothesis in our phylogenetic analyses. The content of the clade Euharamiyida follows previous work<sup>4,38</sup>.

### Phylogenetic analysis

We conducted two sets of phylogenetic analyses separately, using different data matrices to explore the placement of the new taxon in the mammaliaforms and multituberculates. The list of morphological characters for mammaliaform phylogeny follows ref. <sup>4</sup> (derived from refs. <sup>38,39</sup>), with separate analysis of two character matrices, A and B. We created a data matrix for multituberculate phylogeny analysis by adding new taxa and characters to expand the matrix in order to include 51 taxa and 130 characters on the basis of a newly published data matrix<sup>40</sup> (see Supplementary Information). Data matrices were edited in Mesquite v.3.03 and saved in NEXUS format for parsimony and Bayesian analysis. Bayesian analysis for mammaliaform or multituberculate phylogeny was run for 100 million Markov Chain Monte Carlo generations, with the first 25% discarded as ‘burn-in’, using the Mkv model for discrete morphological data and a gamma parameter for rate variation in MrBayes 3.2 (ref. <sup>41</sup>). Posterior probabilities were calculated to assess node robustness in MrBayes. Parsimony analysis was performed using TNT 1.5 with the New Technology Search method, implementing sectorial search, ratchet, drift and tree fusing, under equally weighted parsimony<sup>42</sup>. As is conventional for large datasets, 200 ratchet iterations, 100 drift cycles and 10 rounds of tree fusion were applied to conduct comprehensive searches during phylogenetic analysis. Two separate parsimony analyses were conducted, one with all characters unordered and the other with 19 characters ordered for

the multituberculate data matrix, respectively. These ordered characters are 17, 25, 26, 29, 31, 32, 43, 46, 47, 48, 49, 51, 52, 55, 58, 59, 61, 72 and 85, as suggested previously<sup>43,44</sup>. Node support is given as Bremer support values in strict consensus of parsimony analysis, and as posterior probabilities (percentage) in 50% majority-rule consensus of Bayesian analysis.

### Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

### Data availability

The specimen (IVPP V20778) reported here is housed in the Institute of Vertebrate Paleontology and Paleoanthropology, Beijing, China. Character matrices are given in the Supplementary Information.

- Luo, Z. X., Crompton, A. W. & Sun, A. L. A new mammaliaform from the early Jurassic and evolution of mammalian characteristics. *Science* **292**, 1535–1540 (2001).
- Luo, Z. X., Ji, Q. & Yuan, C. X. Convergent dental adaptations in pseudo-tribosphenic and tribosphenic mammals. *Nature* **450**, 93–97 (2007).
- Rich, T. H. et al. The mandible and dentition of the Early Cretaceous monotreme *Teinolophos trusleri*. *Alcheringa* **40**, 475–501 (2016).
- Luo, Z. X., Gatesy, S. M., Jenkins, F. A. Jr, Amaral, W. W. & Shubin, N. H. Mandibular and dental characteristics of Late Triassic mammaliaform *Haramiyavia* and their ramifications for basal mammal evolution. *Proc. Natl. Acad. Sci. USA* **112**, E7101–E7109 (2015).
- Hu, Y., Wang, Y., Luo, Z. & Li, C. A new symmetrodont mammal from China and its implications for mammalian evolution. *Nature* **390**, 137–142 (1997).
- Lillegraven, J. A. & Krusat, G. Cranio-mandibular anatomy of *Haldanodon exspectatus* (Docodontia; Mammalia) from the Late Jurassic of Portugal and its implications to the evolution of mammalian characters. *Rocky Mountain Geol.* **28**, 39–138 (1991).
- Rowe, T. Definition, diagnosis and origin of Mammalia. *J. Vertebr. Paleontol.* **8**, 241–264 (1988).
- Bi, S., Wang, Y., Guan, J., Sheng, X. & Meng, J. Three new Jurassic euharamiyidan species reinforce early divergence of mammals. *Nature* **514**, 579–584 (2014).
- Krause, D. W. et al. First cranial remains of a gondwanatherian mammal reveal remarkable mosaicism. *Nature* **515**, 512–517 (2014).
- Csiki-Sava, Z., Vremir, M., Meng, J., Brusatte, S. L. & Norell, M. A. Dome-headed, small-brained island mammal from the Late Cretaceous of Romania. *Proc. Natl. Acad. Sci. USA* **115**, 4857–4862 (2018).
- Ronquist, F. et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012).
- Goloboff, P. A., Farris, J. S. & Nixon, K. C. TNT, a free program for phylogenetic analysis. *Cladistics* **24**, 774–786 (2008).
- Yuan, C. X., Ji, Q., Meng, Q. J., Tabrum, A. R. & Luo, Z. X. Earliest evolution of multituberculate mammals revealed by a new Jurassic fossil. *Science* **341**, 779–783 (2013).
- Xu, L. et al. Largest known Mesozoic multituberculate from Eurasia and implications for multituberculate evolution and biology. *Sci. Rep.* **5**, 14950 (2015).

**Acknowledgements** We thank S.-H. Xie for specimen preparation; Y.-M. Hou and P.-F. Yin for help with computed laminography scans and virtual reconstructions; X. Jin and X.-C. Guo for help with photographing and drawing; and T. Martin and J. A. Schultz for access to Guimarota specimens in the University of Bonn. We benefited from discussions with D. W. Krause, Z. X. Luo, T. Martin, J. A. Schultz, N. Kusuhashi and J. K. O’Connor. Financial support was from the Strategic Priority Research Program of the Chinese Academy of Sciences (grants XDB18000000 and XDB26000000), the National Natural Science Foundation of China (grants 41802005 and 41688103), and the State Key Laboratory of Palaeobiology and Stratigraphy (Nanjing Institute of Geology and Palaeontology, CAS; grant 183121).

**Author contributions** Y.W. and H.W. designed the study. H.W. organized computed tomography scans and virtual reconstructions, performed phylogenetic analyses, and prepared the main text, figures and Supplementary Information. All authors contributed to revising the manuscript and figures. Y.W. supervised all research activities.

**Competing interests** The authors declare no competing interests.

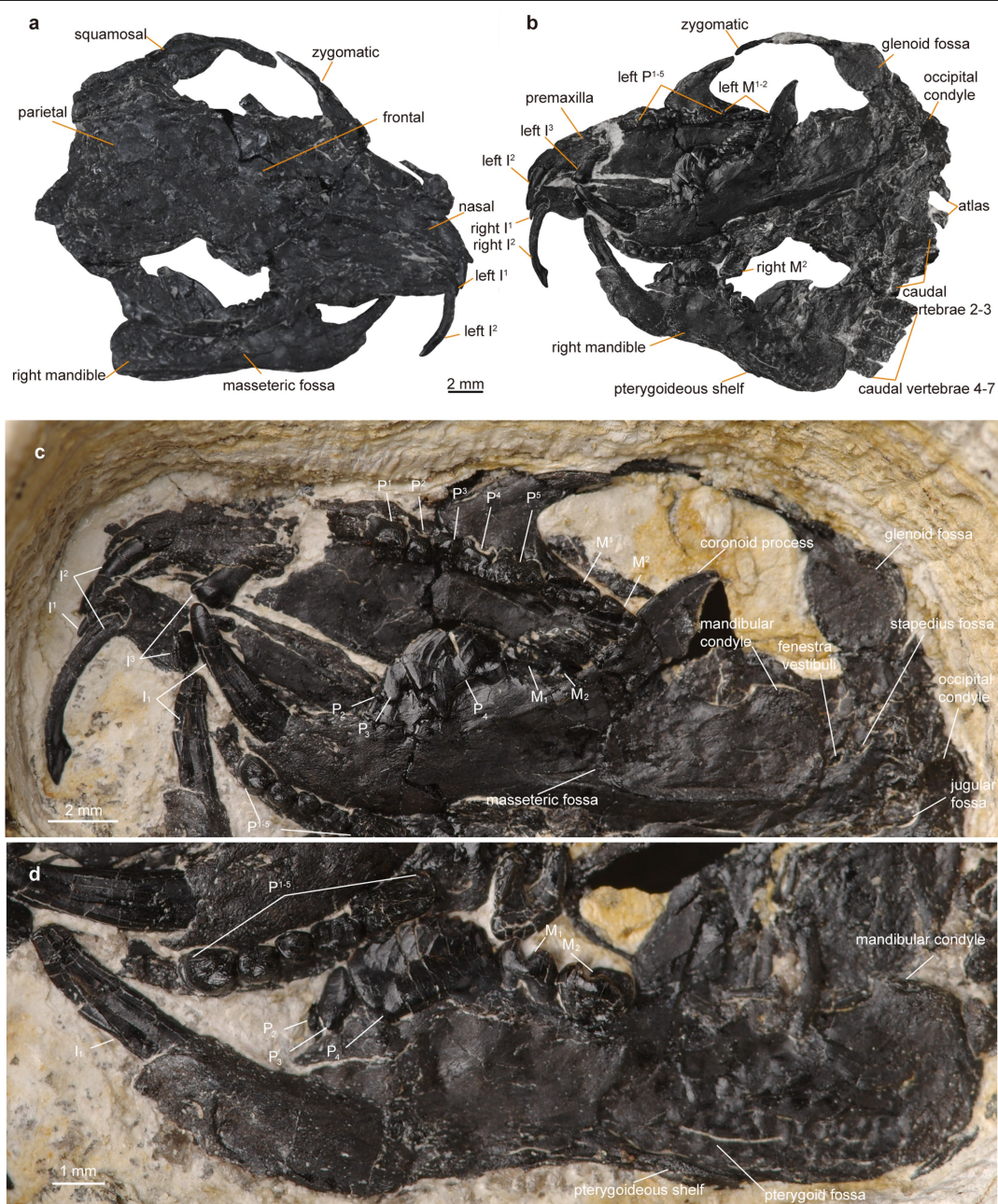
### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1792-0>.

**Correspondence** and requests for materials should be addressed to Y.W.

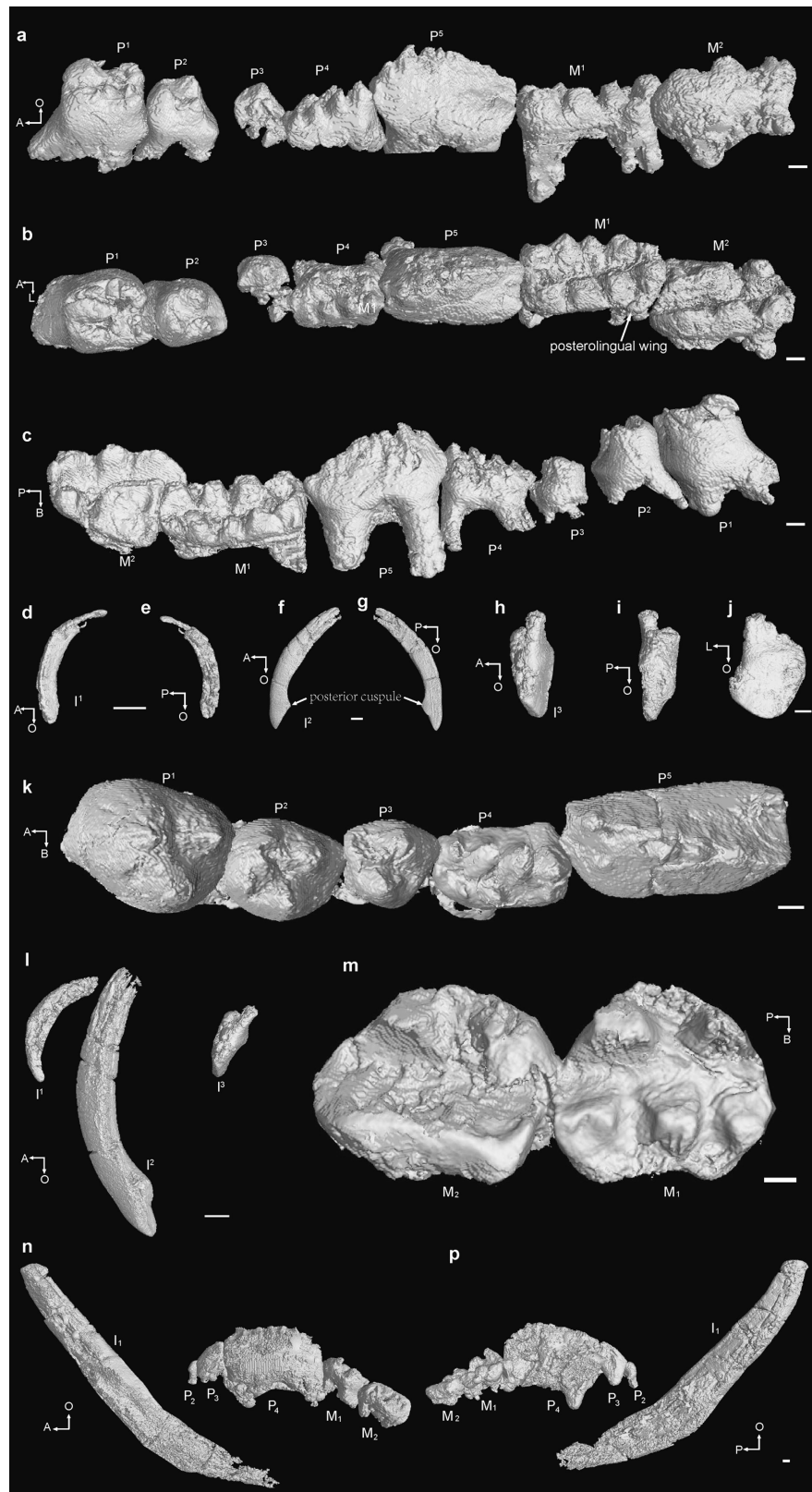
**Peer review information** Nature thanks Simone Hoffmann and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



**Extended Data Fig. 1 | Cranio-mandibular morphology of *J. kielanae* (holotype IVPP V20778).** **a**, Skull in dorsal view and right mandible in lateral view. **b**, Skull in ventral view, left mandible in lateral view, and right mandible in medial view. **c**, Close-up view of cranio-mandibular features. **d**, Close-up medial view of the right dentary. The flat glenoid fossa accommodates the mandibular

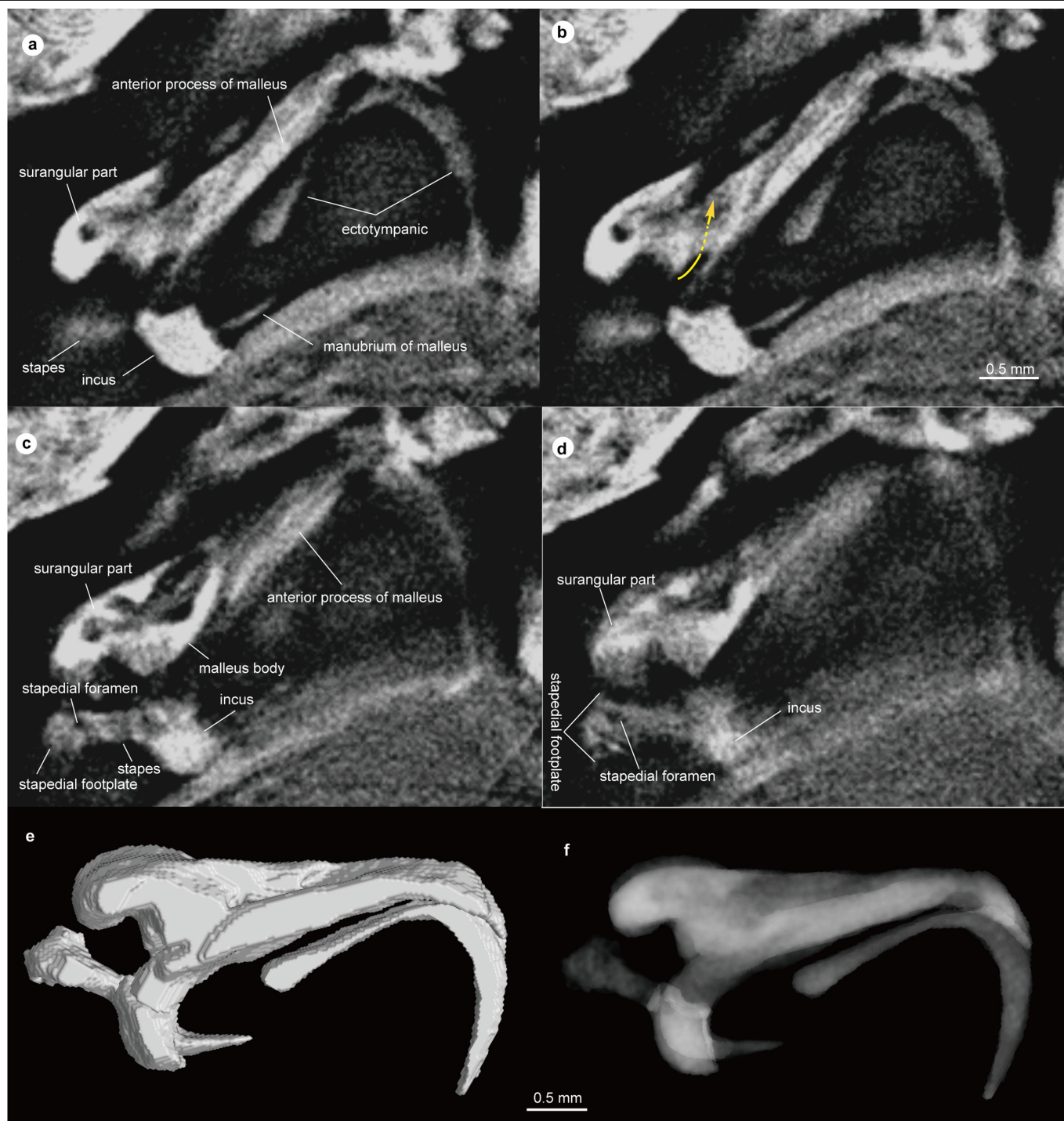
condyle, which is positioned below the occlusal level of the lower molars and faces posteriorly in IVPP V20778. Together with the distinct masseteric fossa—which presumably provides attachment for a well developed masseteric muscle, inserting anteriorly below P<sub>4</sub>—the glenoid fossa produces a palinal (posterior) power stroke with distinct posterior chewing.



**Extended Data Fig. 2 | Dentition of *J. kielanae* (IVPP V20778).** a–c, Left upper cheek teeth ( $P^1$  to  $M^2$ ) in lingual (a), occlusal (b) and buccal (c) views. d, e, Right  $I^1$  in medial view (d) and lateral view (e). f, g, Right  $I^2$  in medial (f) and lateral (g) views. h–j, Right  $I^3$  in lingual (h), buccal (i) and posterior (j) views. k, Right

upper premolars ( $P^1$  to  $P^5$ ) in occlusal view. l, Right upper incisors ( $I^1$  to  $I^3$ ) in medial view. m, Right lower molars ( $M_1$  and  $M_2$ ) in occlusal view. n, p, Right lower teeth ( $I_1$ ,  $P_2$  to  $M_2$ ) in lingual (n) and buccal (p) views. A, anterior; B, buccal; L, lingual; O, occlusal; P, posterior. Scale bars, 0.2 mm.

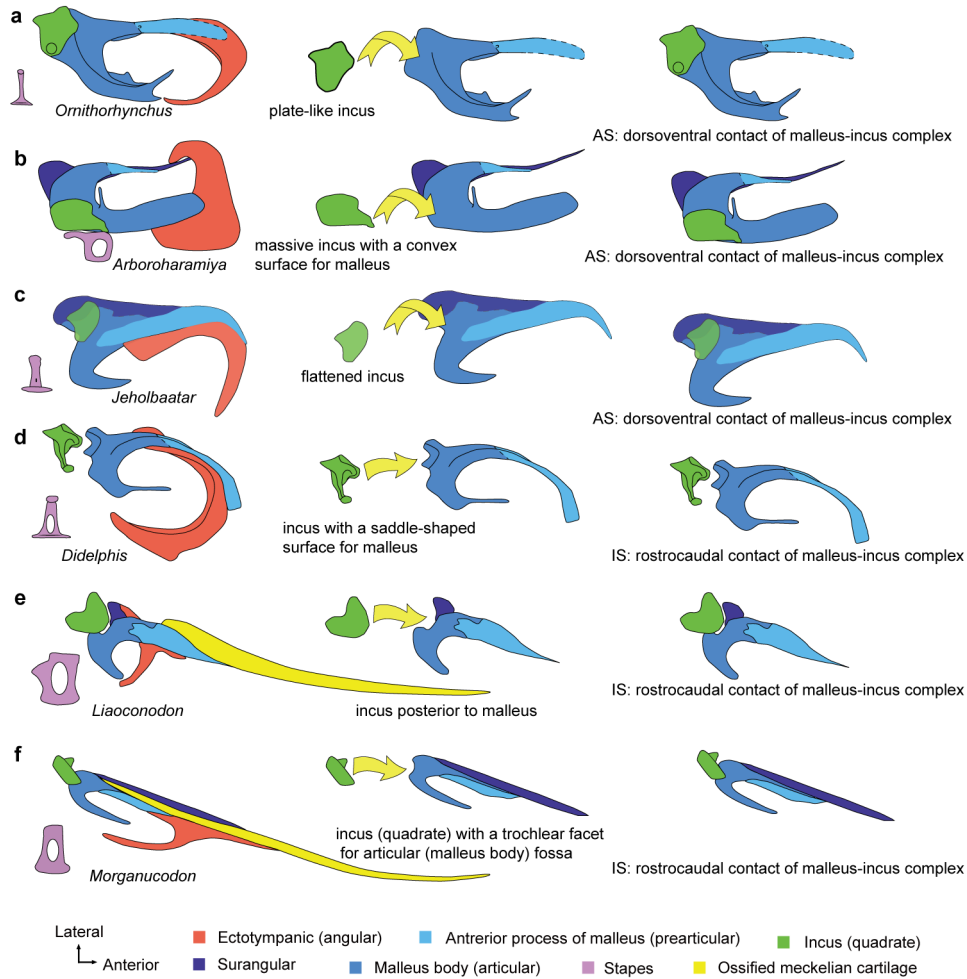




**Extended Data Fig. 3 | Computed laminography images and reconstructions of left middle-ear bones.** **a–d**, Computed laminography images on different levels. The path of the chorda tympani is marked with a yellow arrow in **b**. The stapedial foramen, identified by computed laminography, is shown in **c, d**.

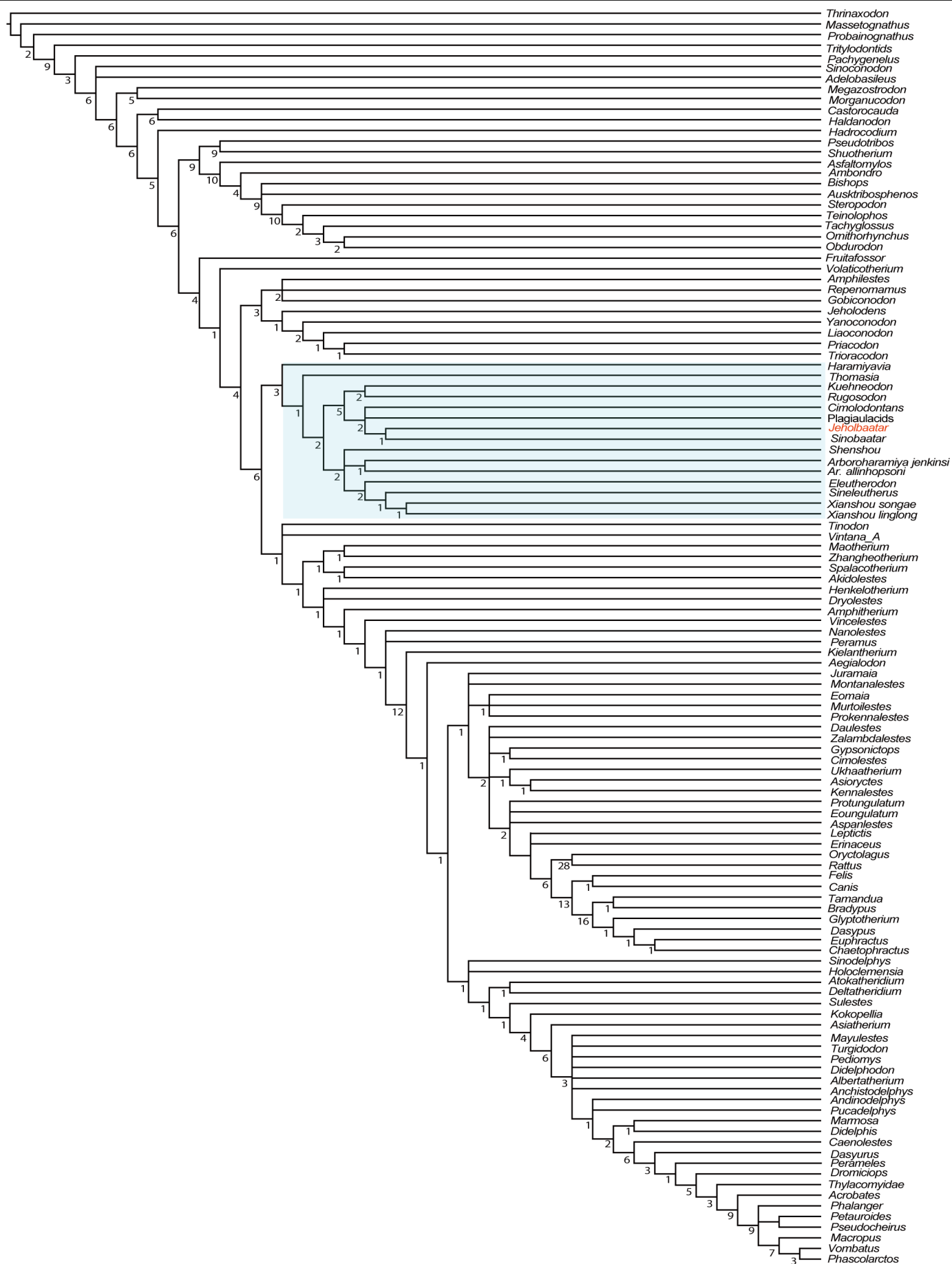
**e**, Three-dimensional reconstruction of left middle-ear bones in dorsal view. **f**, X-ray rendering of left middle ear, showing the differing thicknesses of different parts of the bones.





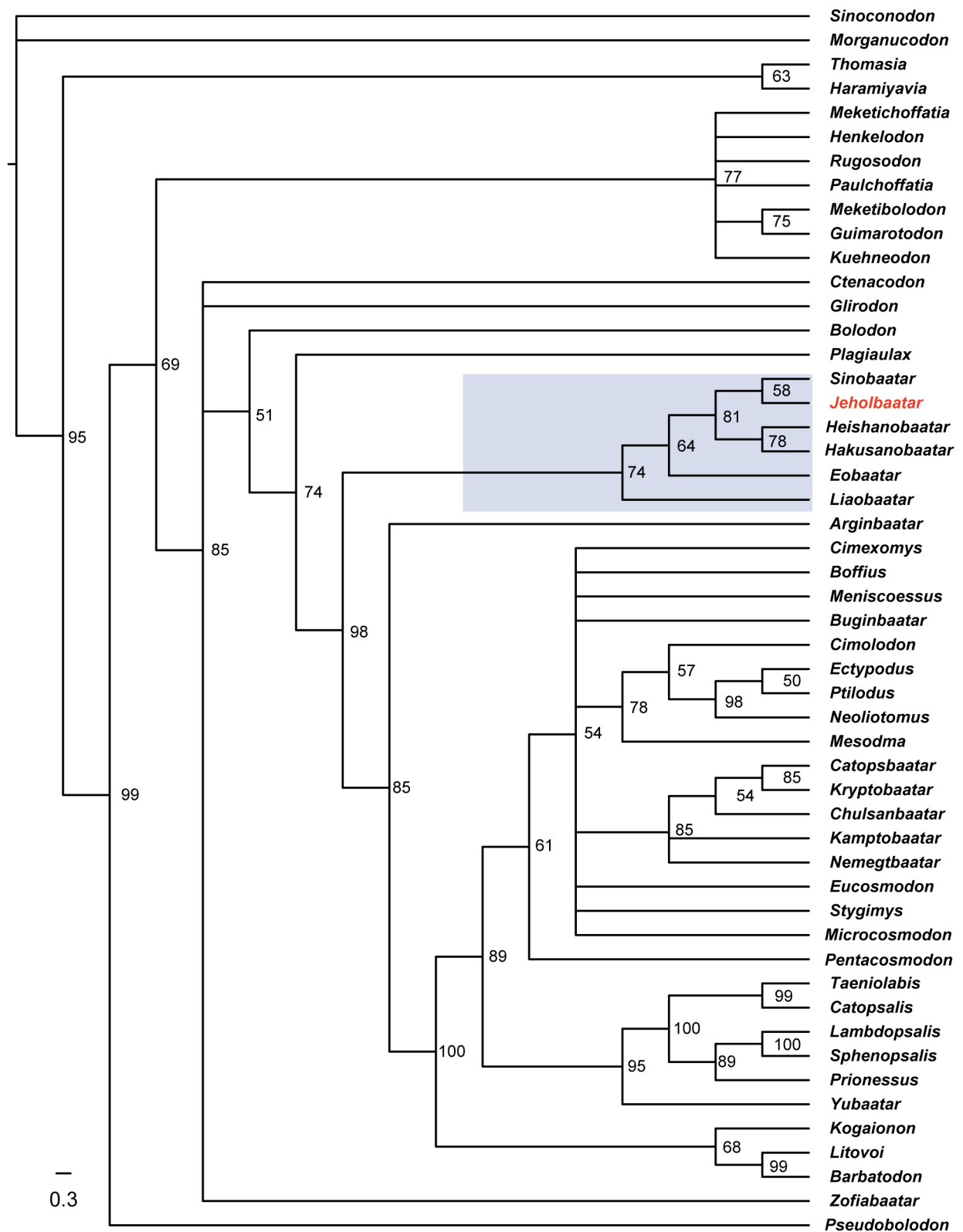
**Extended Data Fig. 4 | Articular configurations of the malleus-incus complex.** **a**, Left auditory bones of *Ornithorhynchus* in dorsal view (modified from ref. <sup>16</sup>). **b**, Interpretive reconstruction of left auditory bones of *Arboroharamiya* in dorsal view (modified from ref. <sup>4</sup>). **c**, Interpretive reconstruction of left auditory bones of *Jeholbaatar* in dorsal view. The yellow arrows in **a–c** show that the incus lies dorsal to the malleus in *Ornithorhynchus*, *Arboroharamiya* and *Jeholbaatar*, demonstrating the ‘abutting system’ (AS) arrangement of the malleus-incus complex. **d**, Left auditory bones of *Didelphis*

in medial view (modified from ref. <sup>28</sup>), showing that the malleus-incus complex maintains the interlocking system (IS) arrangement (yellow arrow), with a rostrocaudal contact between these two elements. **e**, Left auditory bones of *Liaconodon* in medial view (modified from ref. <sup>3</sup>). **f**, Left auditory bones of *Morganucodon* in medial view (modified from ref. <sup>28</sup>). Here the incus (quadrate) has a medial trochlear facet to contact the concave surface of the malleus body (articular fossa) posteriorly.



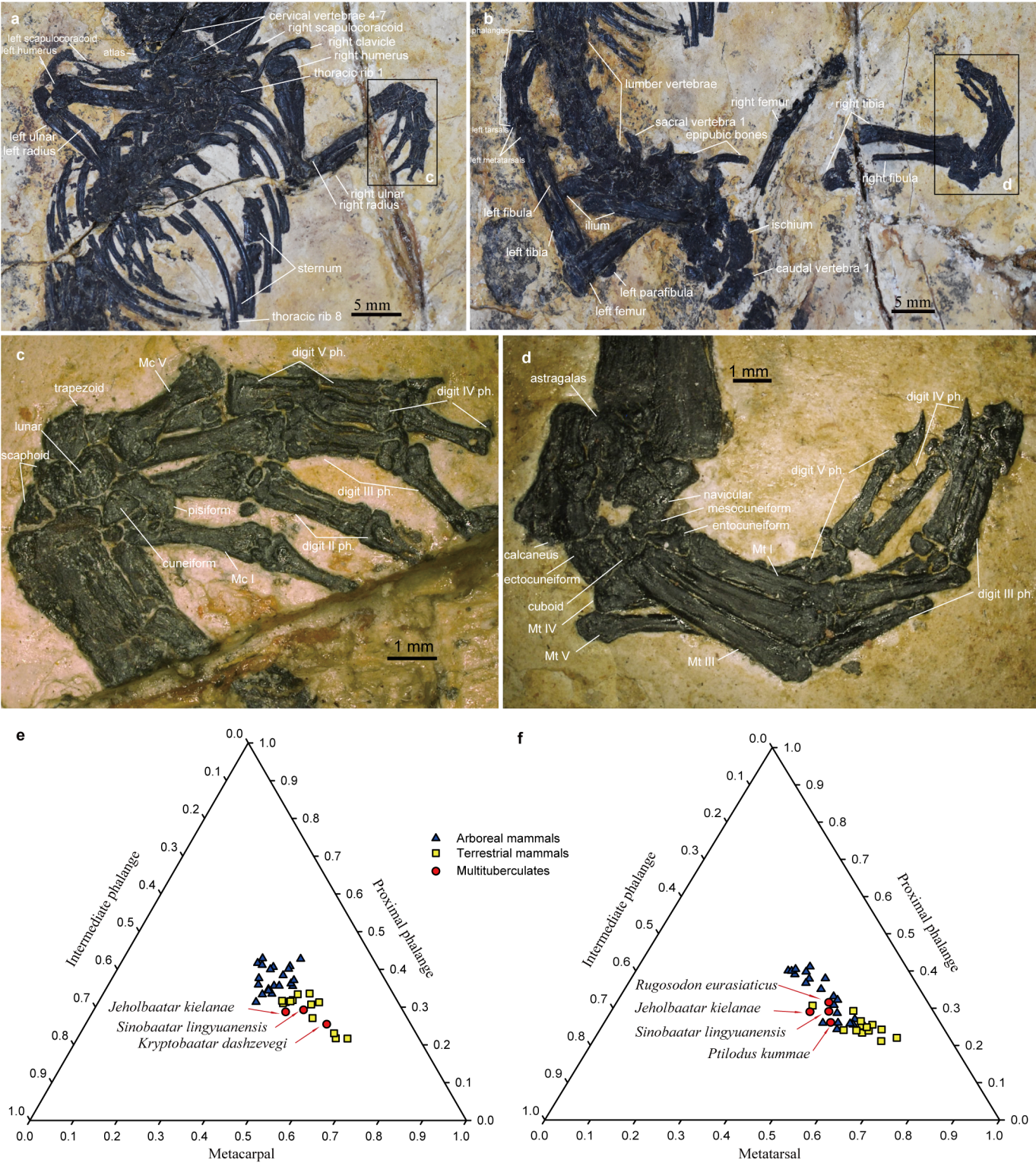
Extended Data Fig. 5 | See next page for caption.

**Extended Data Fig. 5 | Strict consensus of parsimony analysis based on data matrix A.** Tree length, 2,622; consistency index, 0.327; retention index, 0.795. On the basis of analysis using TNT 3.0, 14 most parsimonious trees are returned; tree length, 2,539, consistency index, 0.338; retention index, 0.804. The blue shading shows the monophyly of all otherians within crown mammals. Node supports are given as Bremer support values.



**Extended Data Fig. 6 | Results of Bayesian analysis of multituberculates.**  
This 50% majority-rule consensus was obtained from 10 million Markov Chain Monte Carlo generations with a 25% burn-in fraction. Node supports are listed

as posterior probabilities (percentages). The blue rectangle shows the monophyly of eobaatarids, with *Jeholbaatar* closely related to *Sinobaatar*.



Extended Data Fig. 7 | See next page for caption.



**Extended Data Fig. 7 | Manual and pedal structure, and ternary diagrams showing the proportions of phalanges from manual and pedal digit III.**

**a, b**, Shoulder (**a**) and pelvic (**b**) girdles in dorsal view. **c, d**, Right manus (**c**) and pes (**d**) in lateral view. **e, f**, Ternary plots showing ratios of metapodial (metacarpal or metatarsal), proximal and intermediate phalanges for *Jeholbaatar* digit III from the manus (**e**) and pes (**f**), and comparison with some extant terrestrial and arboreal mammals. The lengths of these three phalanges are shown as ratios of the combined length of these elements. Mc, metacarpal; Mt, metatarsal. The lengths of *Jeholbaatar* manus and pes elements (in mm, with asterisks indicating damaged elements) are: Mc I, 2.76; Mc II, \*2.84; Mc III, \*3.70; Mc IV, \*2.81; Mc V, 2.79; digit I proximal phalanx, 1.98; digit II proximal phalanx, 2.84; digit II intermediate phalanx, \*1.60; digit III proximal phalanx, 2.40; digit III intermediate phalanx, 2.26; digit IV proximal phalanx, \*2.22; digit

IV intermediate phalanx, 1.83; digit V proximal phalanx, 1.92; digit V intermediate phalanx, 1.54; phalange index, that is, (proximal plus intermediate)/metacarpal, digit III, 126%; Mt I, 3.92; Mt II, 4.99; Mt III, 5.42; Mt IV, \*1.69; Mt V, \*3.33; digit I proximal phalanx, 3.51; digit II proximal phalanx, 3.58; digit II intermediate phalanx, 2.82; digit III proximal phalanx, 3.59; digit III intermediate phalanx, 3.46; digit IV proximal phalanx, \*1.73; digit IV intermediate phalanx, 3.25; digit V intermediate phalanx, 2.63; phalanx index, that is, (proximal+intermediate phalanges)/metatarsal, digit III, 130%. The manual proportion of *J. kielanae* places it closer (than the other multituberculates in the sample) to the arboreal category; the pedal proportion clusters mostly with arboreal taxa. The data for extant taxa are derived from ref. <sup>38</sup>.

Extended Data Table 1 | Phalange indices for digit III of *Jeholbaatar* and comparison with other mammals

Taxa	Substrate	Phalange Index	
		Manual digit III	Pedal digit III
<i>Maotherium sinensis</i>	Terrestrial	95%	99%
<i>Didelphis virginiana</i>	Terrestrial	98%	88%
<i>Eomaia scansoria</i>	Scansorial/Terrestrial	130%	129%
<i>Caluromys philander</i>	Scansorial/Terrestrial	138%	156%
<i>Arboroharamiya jenkinsi</i>	Arboreal	246%	216%
<i>Kryptobaatar dashzevegi</i>	Terrestrial	81%	
<i>Rugosodon eurasiatricus</i>	Scansorial/Terrestrial	117%	114%
<i>Sinobaatar lingyuanensis</i>	Scansorial/Terrestrial	108%	109%
<i>Ptilodus kummae</i>	Arboreal	118%	
<i>Eucosmodon</i> sp.	Arboreal	119%	
<i>Jeholbaatar kielanae</i>	Scansorial/Terrestrial	126%	130%

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- |                                     |   |
|-------------------------------------|---|
| n/a                                 | Confirmed   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> The exact sample size ( <i>n</i> ) for each experimental group/condition, given as a discrete number and unit of measurement   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of all covariates tested   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i> ) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted<br><i>Give P values as exact values whenever suitable.</i>                     |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i> ), indicating how they were calculated   |

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection

Data collection for building character matrix used in this study is based on observation of specimens.

Data analysis

Segmentation was conducted in VGStudio v.3.0. Character matrix was compiled in Mesquite v.3.03. For phylogenetic analysis, parsimony analysis was conducted in TNT v. 1.5 and Bayesian analysis was conducted in MrBayes v. 3.2. Measurements were taken in ImageJ.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The specimen reported in this study is housed in an academic institute and available for scholars to examine. Data matrices were provided in Supplementary Information.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	It is a study on only one fossil specimen with phylogenetic analysis.
Research sample	It is a fossil mammal specimen from the Lower Cretaceous of Northeast China. Phylogenetic analysis is based on character matrices that cover various taxa from all mammaliaform clades with emphasis on multituberculates.
Sampling strategy	Fossil collecting in fieldwork and specimens preparation in lab. The taxon sampling and characters selected are extensive enough to reconstruct phylogeny for both mammaliaforms and multituberculates.
Data collection	Data collection includes observation of specimens with microscope in lab and computed laminography.
Timing and spatial scale	H.W. collected data from July, 2015 to June, 2018, based on observations of specimens.
Data exclusions	No data was excluded
Reproducibility	Phylogenetic analysis is repeatable, following the method for both parsimony analysis (in TNT) and Bayesian analysis (in MrBayes) in Method section.
Randomization	N/A. It is a study on fossil material.
Blinding	N/A. Character matrices are built on the basis of independent observation for each taxon.
Did the study involve field work?	<input checked="" type="checkbox"/> Yes <input type="checkbox"/> No

## Field work, collection and transport

Field conditions	Annual average temperature is 5.4°C ~ 8.7°C and annual precipitation is 450-480mm.
Location	Chaoyang City, Liaoning province, China.
Access and import/export	We investigate the fossil locality, prepare the specimen, and scan the specimen at the Institute of Vertebrate Paleontology and Paleoanthropology.
Disturbance	N/A

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input type="checkbox"/>	<input checked="" type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Palaeontology

Specimen provenance	This specimen was discovered from the Jiufotang Formation in Changzigou site, Lingyuan county, Lioaning province, China. No permits needed for the work.
Specimen deposition	The specimen reported in this study is housed in the Institute of Vertebrate Paleontology and Paleoanthropology, Beijing, China.
Dating methods	No new dates for the specimen.

☒ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.



# The GenomeAsia 100K Project enables genetic discoveries across Asia

<https://doi.org/10.1038/s41586-019-1793-z>

GenomeAsia100K Consortium\*

Received: 29 January 2018

Accepted: 11 October 2019

Published online: 4 December 2019

Open access

The underrepresentation of non-Europeans in human genetic studies so far has limited the diversity of individuals in genomic datasets and led to reduced medical relevance for a large proportion of the world's population. Population-specific reference genome datasets as well as genome-wide association studies in diverse populations are needed to address this issue. Here we describe the pilot phase of the GenomeAsia 100K Project. This includes a whole-genome sequencing reference dataset from 1,739 individuals of 219 population groups and 64 countries across Asia. We catalogue genetic variation, population structure, disease associations and founder effects. We also explore the use of this dataset in imputation, to facilitate genetic studies in populations across Asia and worldwide.

The underrepresentation of non-European individuals in human genetic studies<sup>1</sup> limits the applicability of the results for a large proportion of the world's population<sup>2</sup>. Reference genome datasets<sup>3–12</sup> are needed to characterize population-specific variation, enable efficient imputation of variants that are not directly genotyped, and extend genome-wide association studies (GWAS) to additional populations. The value of population-specific reference datasets is well recognized and projects based in the United States and Europe have provided deep characterization of specific populations (for example, Ashkenazi Jews<sup>12</sup> and individuals from the Netherlands<sup>3</sup> and Iceland<sup>13</sup>) and, in particular, data from individuals of Nordic countries have provided examples of how reference genome datasets can be used to drive comprehensive genetic studies across an entire population<sup>14</sup>. In Africa, populations show complex genetic patterns, smaller blocks of linkage disequilibrium and higher levels of heterozygosity, which provides unique value for genetic studies. Across the continent, early reference genome datasets for diverse populations are being built as part of H3Africa and other studies<sup>5,15</sup>. A Korean reference genome as well as Japanese and Chinese reference genome datasets have been created, and the formation of large biobanks such as BioBank Japan<sup>16</sup> and the China Kadoorie Biobank<sup>17</sup> will accelerate the pace of discovery of disease associations across east Asia.

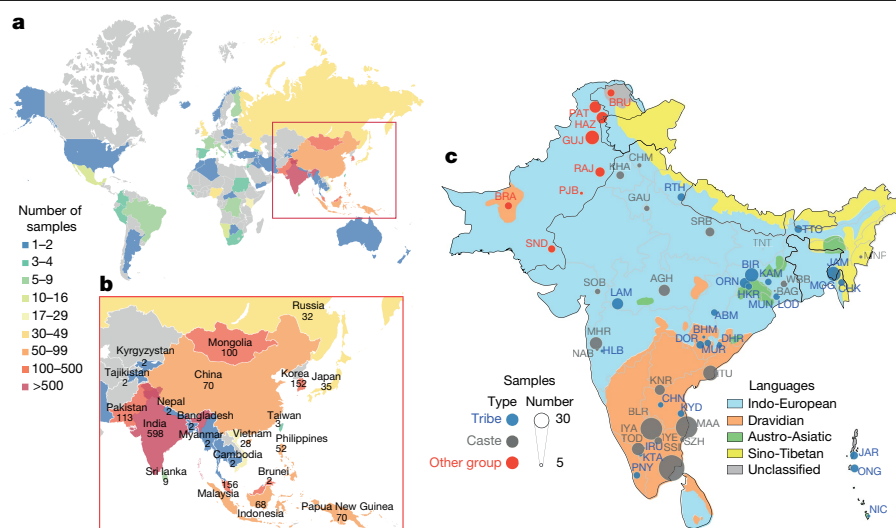
A shared recognition of the value of coordinated efforts and the need for reference genome datasets that would be useful for the complex populations of Asia has led to the formation of the GenomeAsia consortium (<http://www.genomeasia100k.com>). The consortium serves to facilitate and coordinate sequencing efforts among consortium members to maximize the value of the genomic sequence data that is produced and to facilitate efforts by national or other regional groups. Here we describe the GenomeAsia Pilot (GAsP) project, which consists of analyses of the whole-genome sequencing data of 1,739 individuals from 219 population groups across Asia, with the ultimate goal of providing a useful genomic resource and facilitating genetic studies in Asia. We use the data that was generated in this pilot to analyse population structure and history, and as the basis for designing larger-scale genomic studies. Furthermore, we explore disease-associated loci as an initial comparison of differences between populations. We show that

the variant data produced by this project improve variant filtering for the discovery of disease-associated genes of rare diseases. We show that Asia has sizable founder populations and that further studies in these populations may be useful for the discovery of rare-disease-associated genes. We also report an initial survey of loss-of-function alleles found in the GAsP project.

## The GAsP dataset

For the GAsP project, we generated 1,267 high-coverage (average 36×) whole-genome sequences and analysed these together with 596 publicly available human genome sequences from previous sequencing studies (Supplementary Information 1, 2 and Supplementary Tables 1a–c, 2a). The 1,739 samples were enriched for individuals from population isolates to capture the broadest wealth of genetic diversity; the dataset includes 598 sequences from India, 156 from Malaysia, 152 from South Korea, 113 from Pakistan, 100 from Mongolia, 70 from China, 70 from Papua New Guinea, 68 from Indonesia, 52 from the Philippines, 35 from Japan and 32 from Russia (Fig. 1a–c and Supplementary Table 1a–c). To facilitate comprehensive and comparative analysis of human genetic variation, we included sequencing data from African, European and American samples (Supplementary Table 1a, b). The sequenced samples originate from 7 global regions, 64 different countries of origin and 219 population groups. About 80% of the samples come from Asia and emphasize population groups that are underrepresented in previous genetic studies (Fig. 1a, b, Supplementary Tables 1a–c, 2b and Supplementary Information 1, 2). Each global region and population group was assigned a unique three-letter code for future reference (see Supplementary Table 1a for three-letter code designations). Within Asia, the sampling of many distinct population groups allowed us to analyse the relationship between geography, physical characteristics and genetic variation. In south and southeast Asia, in particular, we sampled across diverse populations to gather new insights into how groupings defined on the basis of caste and language relate to genetic diversity, admixture with extinct hominins and other genetically described characteristics.

\*A list of participants and their affiliations appears in the online version of the paper.



**Fig. 1 | Sampling distribution of GASP.**

**a, b**, Sample sizes. **c**, Location, language and social hierarchy associated with samples from south Asia. Groups with fewer than three samples are not plotted. See Supplementary Table 1a for definitions and descriptions of samples and population groups included in each geographically defined set.

## Population structure

Knowledge of the complex history of Asian populations informs optimal sampling for larger-scale biomedical sequencing efforts. We applied standard approaches for detecting recent positive selection, quantifying the population structure and inferring the history of the different populations, including principal component analysis<sup>18</sup>, multiple sequentially Markovian coalescent (MSMC)<sup>19</sup>, ADMIXTURE<sup>20</sup>,  $F_{ST}$ , uniparental analyses and the analysis of the Y chromosome and mitochondrial haplogroups (Fig. 2, Extended Data Fig. 1 and Supplementary Information 3–10). Our results generally recapitulate the broad inferences of previous studies, and ADMIXTURE plots show complex structure within south and southeast Asia (Fig. 2a). In particular, India, Malaysia and Indonesia contain multiple ancestral populations as well as multiple admixed groups. On the basis of MSMC cross-coalescence rates, which reflect the increase in coalescence times of haplotypes sampled from different populations relative to haplotypes sampled from the same population<sup>19</sup>, we estimate that the oldest population splits in southeast Asia and Oceania involve Melanesians and/or Negritos, who show a substructure from approximately 40 thousand years ago and evidence of separation around 20–30 thousand years ago (Extended Data Fig. 1b and Supplementary Information 3). The population structure provides genetic information on classically defined population groups to aid future studies. For example, using multiple analytical approaches (Supplementary Information 3, 6), we confirmed that the anthropologically classified ‘Negrito’ groups from India, Malaysia and the Philippines, are genetically more closely related to their geographical neighbours than they are to other Negrito groups<sup>21,22</sup>, suggesting that dark skin colour is probably an environmental adaptation (for example, to high levels of solar radiation) and not an indicator of shared ancestry.

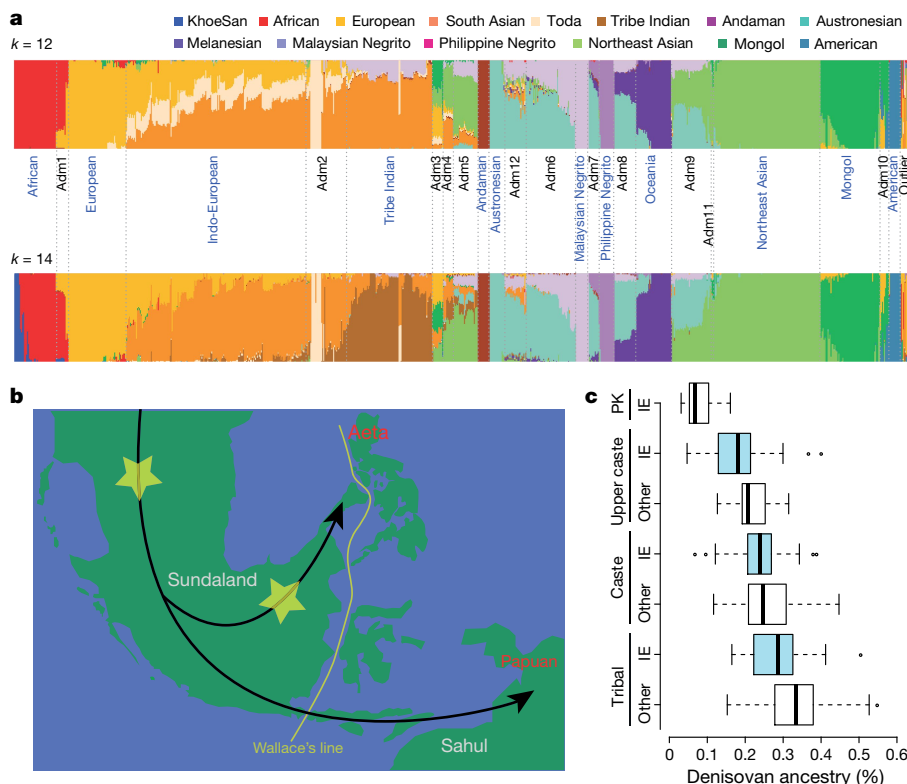
Our dense sampling of Asian populations enables the examination of Denisovan admixture in greater detail than has been previously possible, providing information about population splits or in-flows that occurred at or after the time of admixture (Supplementary Information 10). Our estimates of Denisovan ancestry were highest in Melanesians and the Aeta, intermediate in the Ati and groups from the Indonesian island of Flores, and low (but still significantly greater than 0) in most south, east and southeast Asian populations. We found high levels of Denisovan ancestry in Philippine Negrito groups but not in Malay or Andaman Negritos; these results are qualitatively similar to what was found in a previous study that was based on single-nucleotide polymorphism (SNP) arrays<sup>23</sup>. The high levels of Denisovan ancestry in Melanesians and the Aeta are consistent with an admixture event into a population that is ancestral to both<sup>23</sup>; however, two lines of evidence suggest that the ancestors of the Aeta experienced a second

Denisovan admixture event. First, multiple analyses found that the Aeta are genetically more similar to populations without appreciable Denisovan ancestry (for example, Igorot, Malay and Malay Negrito groups) than they are to Melanesians (Supplementary Information 3, 6). This can be explained by more recent gene flow from other populations without Denisovan ancestry. However, such gene flow would reduce the levels of Denisovan admixture below that found in Melanesians. More directly, we find that putative Denisovan haplotypes that are unique to the Aeta ( $n = 962$ ) are significantly longer than putative Denisovan haplotypes shared between Aeta and Papuans ( $n = 596$ , mean = 16.1 kb compared with mean = 14.1 kb, Mann–Whitney  $U$ -test,  $P < 10^{-10}$ ), or putative Denisovan haplotypes unique to Papuans ( $n = 727$ , mean = 16.1 kb compared with mean = 14.9 kb, Mann–Whitney  $U$ -test,  $P < 10^{-1,000}$ ) (Supplementary Information 10), supporting a scenario in which a second admixture event between the Aeta and Denisovans happened after the separation of the Aeta and Melanesians. Two distinct Denisovan admixture events are most consistent with *Homo sapiens* and Denisovans interacting within southeast Asia<sup>23</sup>, making it likely that admixture occurred within Sundaland (Fig. 2b) or even farther east<sup>24,25</sup>.

A recent study found a slightly increased amount of Denisovan ancestry in south Asians compared with a priori expectations<sup>26</sup>. We examined whether this was correlated with either language or social and/or caste status. South Asian samples were grouped into individuals who speak Indo-European languages and individuals who speak non-Indo-European languages (excluding individuals who speak Tibeto-Burman languages), as well as four social or cultural groups: tribal (Adivasi) groups, lower-caste groups, high-caste groups and Pakistani groups (Indo-European language speaking only). We found that the average levels of Denisovan ancestry were significantly different between the four social or cultural groups (Mann–Whitney  $U$ -test,  $P < 10^{-8}$  for all pairwise comparisons; Fig. 2c and Supplementary Information 10). Our results are consistent with the scenario that Indo-European-speaking migrants who entered the subcontinent from the northwest admixed with an indigenous South Asian (ancestral south Indian)<sup>27,28</sup> group who had higher levels of Denisovan ancestry.

## Medical relevance

We evaluated the use of GASP dataset in disease-associated genetic studies and medically relevant applications to determine how the results of larger continuing GenomeAsia studies can be used to improve human health (Supplementary Table 4a). We annotated high-quality variants using public databases including ExAC (Exome Aggregation Consortium)<sup>29</sup>, gnomAD<sup>29</sup>, 1000 Genomes Project<sup>4</sup>, ESP (NHLBI GO Exome Sequencing Project)<sup>30</sup> and dbSNP (Extended Data Fig. 2) and focused



**Fig. 2 | Population structure and admixture.**

**a**, ADMIXTURE plots for  $k=12$  and  $k=14$  illustrating the identification of 12 reference groups.

**b**, Proposed modern human migration route into southeast Asia during the Last Glacial Maximum with potential locations of Denisovan admixture (yellow asterisks). Green indicates the above water landmass at the glacial maximum and white outlines indicate present-day shorelines.

**c**, Estimates of Denisovan ancestry in south Asians, stratified by social/cultural group and language. IE, Indo-European. Adivasi Indo-European,  $n=30$ ; Adivasi non-Indo-European,  $n=196$ ; caste Indo-European,  $n=68$ ; caste non-Indo-European,  $n=155$ ; upper caste Indo-European,  $n=49$ ; upper caste non-Indo-European,  $n=19$ ; Pakistani Indo-European,  $n=79$ . The centre line indicates the median; box limits show the middle 50%; whiskers extend two standard deviations from the mean; points are outliers.

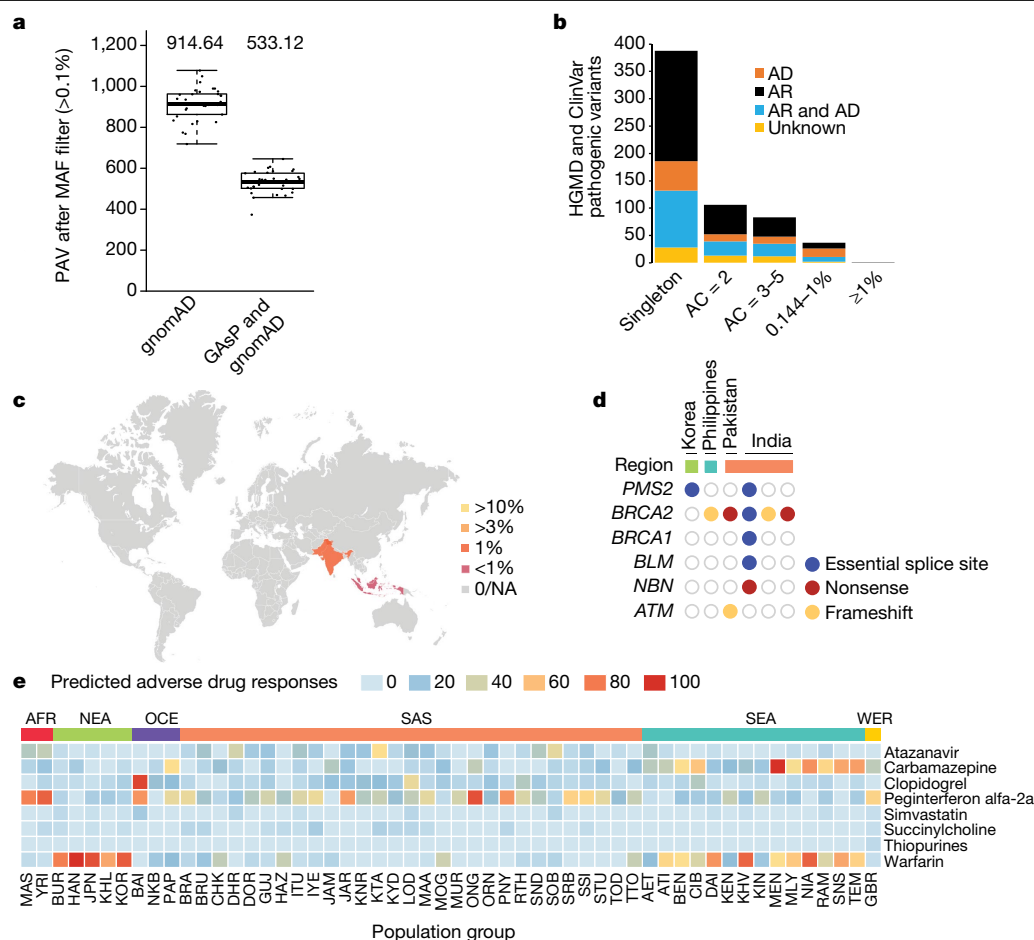
on coding-sequence variants. Overall 23% of protein-altering variants in GASp were not found in these data sources. As expected the majority of coding variants were singletons or very rare (Extended Data Fig. 2). However, the absolute numbers of novel variants with a minor allele frequency (MAF)  $\geq 0.1\%$  within our pan-Asian dataset is large ( $n=194,585$ ), and these are frequent enough to be of relevance for large-scale genetic association studies. We also searched for variants present at low frequency in the overall dataset that are present at significantly higher allele frequencies in one or more of the population groups. We found an additional 144,329 novel variants with MAF  $> 1\%$  in the full GASp dataset that were present at a frequency of greater than 1% within populations grouped by geography; South Asia, Southeast Asia, Northeast Asia or Oceania (see Supplementary Table 1a for description of samples and population groups included in each geographically defined set). These geographical regions contain many diverse population groups, and additional studies are needed to characterize patterns of genetic variation in these groups and disease relevance.

In rare disease genetics, databases are used to filter based on allele frequency with the idea that common alleles are unlikely to be responsible for rare highly penetrant disorders; however, in the absence of appropriate population reference datasets, allele frequencies can be misclassified and may lead to false disease associations<sup>31</sup>. We explored whether the GASp variant dataset can improve the ability to identify disease-relevant variants in Asian cohorts. We analysed 152 exomes from individuals participating in the Indian Maturity Onset Diabetes in the Young (MODY) project. When both the gnomAD and GASp datasets were used for filtering (MAF  $> 0.1\%$ ), we reduced the set of remaining candidate variants by approximately twofold in comparison to using the gnomAD dataset alone (Fig. 3a). In this process, we identified a common population polymorphism in *NEUROD1* (H241Q) that is probably benign but that was previously reported to be medically relevant<sup>32,33</sup>. We annotated variants that were identified in the GASp dataset against the Human Gene Mutation Database (HGMD) disease-causing pathological and ClinVar pathogenic variants. This analysis identified 732 variants (686 SNPs and 46 insertions or deletions (indels)) in 514 genes (Fig. 3b, Supplementary Table 4b, c and Supplementary Information 11). We

compared the 732 pathogenic variants against the gnomAD, ExAC<sup>29</sup>, 1000Genomes<sup>4</sup>, ESP<sup>30</sup>, dbSNP<sup>34</sup>, ALSPAC, TwinsUK<sup>35</sup> and 1000Japanese<sup>6</sup> databases to remove variants that occurred at  $> 1\%$ , focused on those with allele frequencies  $> 0.15\%$  in GASp (38 variants), and reviewed them against the criteria defined by the American College of Medical Genetics (ACMG). This resulted in reclassification of 11 of the 38 variants (Supplementary Table 4d). We examined the geographical distribution of the remaining, revalidated but high-frequency, pathogenic disease-associated variants. As expected, most of these variants were highly enriched in Asia. For example, an HBB variant (chromosome 11: 5248155 c.92+5G>C) associated with  $\beta$ -thalassaemia is found almost exclusively in south Asians and at a lower frequency in southeast Asians (Fig. 3c).

We also examined our dataset for novel variants in genes known to be associated with cancer risk. We found 13 unique variants in 6 genes from 17 samples. This included frameshift, stop-gained and essential splice-site mutations in *BRCA2* ( $n=9$ ), *BRCA1* ( $n=1$ ), *ATM* ( $n=2$ ), *BLM* ( $n=1$ ), *NBN* ( $n=2$ ) and *PMS2* ( $n=2$ ) (Fig. 3d and Supplementary Table 4e). Of the two *PMS2* essential splice variants, one was found in a Korean sample. Loss-of-function mutations in *PMS2* are associated with mismatch repair defects that lead to a higher risk of cancer development. In a separate study of gall bladder cancer, we found the same essential splice site *PMS2* mutation (chromosome 7:6043690C>G) in a Korean patient whose gall bladder cancer exhibits microsatellite instability (E.W.S. and S. Seshagiri, manuscript in preparation). Identification of genetic variants that affect drug efficacy and safety through the alteration of pharmacokinetics enables application of individualized treatment<sup>36–41</sup>. Variation in drug responses are generally recognized and recommendations for dosing are sometimes guided by apparent or self-reported population identity despite the lack of a rigorous pharmacogenomic basis. We assessed the allele frequencies of key pharmacogenomic variants in our dataset to identify inter-population differences that have potential implications on drug testing and treatment (Fig. 3e, Supplementary Table 4g and Supplementary Information 13).

Carbamazepine, clopidogrel, peginterferon and warfarin showed the largest variation between populations in predicted adverse drug responses with groups ranging from 0 and 100 predicted adverse drug



**Fig. 3 | Disease-relevant variant discovery.** **a**, Filtering using the GAsP dataset improves candidate variant discovery by removing population specific variants ( $n = 152$ ). The centre line indicates the median; box limits show the upper and lower quartiles; whiskers extend  $1.5 \times$  the interquartile range. **b**, Allele count (AC) and frequency distribution of variants in the GAsP dataset that are designated disease-causing in the Human Gene Mutation Database (HGMD) or pathogenic in ClinVar. Autosomal-dominant (AD) or autosomal-recessive (AR) or other (unknown) classification as per OMIM. A number of variants ( $n = 37$ ) that had previously been reported to be pathogenic are found

in the GAsP study dataset at high frequency and were reclassified (Supplementary Table 4d). **c**, Frequency of  $\beta$ -thalassaemia variant (chromosome 11:5248155 c.92+5G>C) across Asia shows a geographical enrichment. MAF in South Asia is 1.4%. NA, not available. **d**, Novel cancer-predisposing variants identified in the GenomeAsia dataset. **e**, Population-specific probabilities of adverse drug reactions predicted from the aggregate allele frequencies of known variants associated with response to the indicated drugs.

responses. For example, the HLA-B\*15:02 variant, associated with risk for development of Steven Johnson syndrome<sup>38</sup> in patients treated with carbamazepine was found to occur at an increased frequency in Austronesian language-speaking populations from southeast Asia (for example, 63% in the Mentawai of West Sumatra; 46.6% in the Nias of North Sumatra) compared with other groups (Supplementary Information 13). There are roughly 400 million individuals who belong to Austronesian groups that are at increased risk for carbamazepine sensitivity, including the vast majority of the people from Indonesia, Malaysia and the Philippines.

## Founder populations

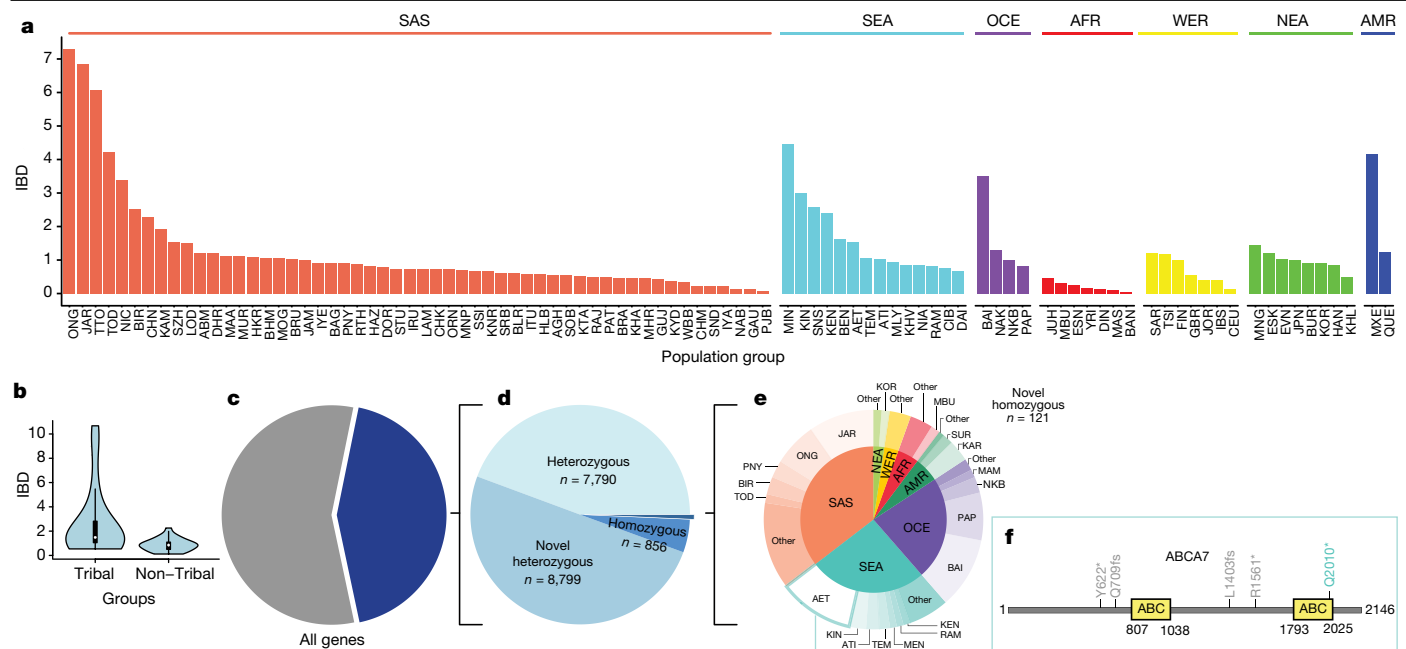
Population bottlenecks produce strong founder effects and increased rates of recessive disease. In populations with strong founder effects, the loss-of-function variant frequency spectrum is skewed higher, greatly increasing power of association<sup>42</sup> and providing unique advantages for the identification of genes associated with both rare and complex diseases<sup>43,44</sup>. We followed the approach described in a previous study on south Asian populations to characterize the degree to which genomic segments are inherited as identical by descent (IBD) in population groups in our dataset<sup>45</sup>.

Our analysis revealed IBD scores of 1.465 and 0.817 for Finnish and British groups, consistent with previous analyses<sup>45</sup>. The IBD score of all of the groups was normalized relative to the Finnish group (Fig. 4a and Supplementary Information 12). Our study includes many groups with small population sizes and it is expected that endogamy paired with small population size will greatly increase IBD scores. We found that indigenous and tribal groups had IBD scores that were skewed upwards from non-tribal groups (Fig. 4b). Notably, we found that a number of Asian groups with large urban populations have IBD scores above or close to that of the Finnish population. For example, samples from an outpatient hospital in Chennai, a city with a census size of 9 million, had an IBD score that was approximately 1.3 times greater than the score for the Finnish group.

## Human knockouts

Homozygous loss-of-function alleles found in humans give us the opportunity to assess the phenotypic effect of specific gene loss and can provide important information about opportunities for treating disease<sup>46,47</sup>. To assess the contents of our dataset, we examined high-confidence protein-truncating variants (PTVs). We found 17,566 PTVs with at least 1 PTV in approximately 43% of all protein-coding genes ( $n = 8,766$ ; Fig. 4c). Among the PTVs, most were heterozygous variants





**Fig. 4 | Founder effects and homozygous loss of function.** **a**, IBD scores across different population groups are shown for 96 ethnicities (1,417 samples) across global regions. The scores given in the figure are relative ratios compared to that of the Finnish group. **b**, Violin plot showing IBD scores in 29 tribal groups and 25 non-tribal groups consisting of 293 and 336 samples, respectively. The centre line indicates the median; box limits show 1.5× the interquartile range.

unique to our dataset ( $n = 8,799$ ; Fig. 4d), similar to the PTV data from ExAC<sup>25</sup> (67% singletons). A smaller number were homozygous and had been reported in gnomAD, dbSNP or 1000 Genomes Project ( $n = 856$ ). In addition, within our dataset were 121 homozygous PTVs that have not previously been reported (Supplementary Table 5). These novel homozygous PTVs were mostly found in groups with high IBD scores such as the Jarawa and Onge from the Andaman Islands (Fig. 4e). The novel homozygous PTVs include an allele of the *ABCA7* gene, Q2010\*, that is found in only the Aeta population (Fig. 4f). Heterozygosity for loss-of-function alleles of *ABCA7* has been shown to increase susceptibility to Alzheimer's disease in European populations<sup>48</sup>.

## Imputation panel

We carried out preliminary work to evaluate the utility of the pilot dataset for imputation. For this analysis, we downsampled whole-genome sequence data from South Asian, Southeast Asian and Northeast Asian population groups (see Supplementary Table 1a for samples included in each of these geographically defined sets) 30× to the genotypes represented on the Illumina Global Screening Array v.1 genotyping array, and compared the imputation using either phase 3 of the 1000 Genomes Project or the GASP reference panels. We found, as described by Illumina, that imputation accuracy of the 1000 Genomes Project reference panel is consistently well below 90% for east Asian and south Asian samples whereas using the GASP reference panel we achieved accuracies ranging from 93 to 95%. To accelerate evaluation and broad utility, we have placed the data on the Michigan Imputation Server (<https://imputationserver.sph.umich.edu/index.html>).

## Discussion

Understanding the genetic basis of human disease will benefit from an increase in the number and scale of disease-association studies that are carried out in Asian populations. In the pilot phase of the GenomeAsia project, the sample set that we analysed allowed us to address a wide range of questions regarding the history of specific Asian population

groups and to map out strategies for additional sequencing efforts. We plan for a staged and coordinated approach, to include the generation of genomic population-specific reference datasets and imputation panels, and use this approach for the production of custom SNP arrays as a catalyst for disease-association studies. This approach is particularly useful in founder populations, such as recent studies in the founder populations of Finland<sup>49</sup>, as well as other populations. This will be particularly valuable in Asia<sup>14,50</sup>, which has founder effects that have not only previously been demonstrated in isolated populations, but are also evident in major urban centres.

Analysis of the GASP dataset allows us to map out strategies for efforts focused on specific population centres in Asia as well as the generation of important tools that will increase our understanding of how genetic variants affect disease susceptibility and drug responses. The dataset improves the ability to filter out low-probability candidates for highly penetrant disorders, to identify putatively pathogenic variants that are found at high frequency in particular populations and improve the ability to infer pathogenicity of identified variants. The identification of novel homozygous PTVs in this study expands the catalogue of genes in which homozygous loss of function appears to be tolerated and, when combined with phenotype information, this will provide important biological insights into gene function. The ability to define gene function in humans through the study of the phenotypic effects of loss-of-function mutations is becoming an increasingly valuable approach<sup>51</sup> and the study of additional variants and populations in which homozygosity occurs at high rates will add to the global resources for carrying out human knockout studies.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1793-z>.



1. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
2. Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
3. The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46**, 818–825 (2014).
4. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
5. Gurdasani, D. et al. The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**, 327–332 (2015).
6. Nagasaki, M. et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat. Commun.* **6**, 8018 (2015).
7. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
8. Mallick, S. et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
9. Mitt, M. et al. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *Eur. J. Hum. Genet.* **25**, 869–876 (2017).
10. Southam, L. et al. Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits. *Nat. Commun.* **8**, 15606 (2017).
11. Xue, Y. et al. Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. *Nat. Commun.* **8**, 15927 (2017).
12. Lencz, T. et al. High-depth whole genome sequencing of an Ashkenazi Jewish reference panel: enhancing sensitivity, accuracy, and imputation. *Hum. Genet.* **137**, 343–355 (2018).
13. Ebenesersdóttir, S. S. et al. Ancient genomes from Iceland reveal the making of a human population. *Science* **360**, 1028–1032 (2018).
14. Njølstad, P. R. et al. Roadmap for a precision-medicine initiative in the Nordic region. *Nat. Genet.* **51**, 924–930 (2019).
15. Bentley, A. R., Callier, S. & Rotimi, C. The emergence of genomic research in Africa and new frameworks for equity in biomedical research. *Ethn. Dis.* **29**, 179–186 (2019).
16. Nagai, A. et al. Overview of the BioBank Japan Project: study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
17. Chen, Z. et al. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int. J. Epidemiol.* **40**, 1652–1666 (2011).
18. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
19. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
20. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
21. The HUGO Pan-Asian SNP Consortium. Mapping human genetic diversity in Asia. *Science* **326**, 1541–1545 (2009).
22. Aghakhanian, F. et al. Unravelling the genetic history of Negritos and indigenous populations of Southeast Asia. *Genome Biol. Evol.* **7**, 1206–1215 (2015).
23. Reich, D. et al. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am. J. Hum. Genet.* **89**, 516–528 (2011).
24. Mijares, A. S. B. The early Austronesian migration to Luzon: perspectives from the Peñablanca cave sites. *Bull. Indo-Pacific Prehist. Assoc.* **26**, 72–78 (2006).
25. Détroit, F. et al. A new species of *Homo* from the Late Pleistocene of the Philippines. *Nature* **568**, 181–186 (2019).
26. Sankararaman, S., Mallick, S., Patterson, N. & Reich, D. The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. *Curr. Biol.* **26**, 1241–1247 (2016).
27. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489–494 (2009).
28. Majumder, P. P. & Basu, A. A genomic view of the peopling and population structure of India. *Cold Spring Harb. Perspect. Biol.* **7**, a008540 (2015).
29. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
30. NHLBI GO Exome Sequencing Project (ESP). *Exome Variant Server*. <http://evs.washington.edu/EVS/> (version: ESP6500SI-V2) (2015).
31. Piton, A., Redin, C. & Mandel, J. L. XLID-causing mutations and associated genes challenged in light of data from large-scale human exome sequencing. *Am. J. Hum. Genet.* **93**, 368–383 (2013).
32. Chapla, A. et al. Maturity onset diabetes of the young in India - a distinctive mutation pattern identified through targeted next-generation sequencing. *Clin. Endocrinol.* **82**, 533–542 (2015).
33. Mohan, V. et al. Comprehensive genomic analysis identifies pathogenic variants in Maturity-Onset Diabetes of the Young (MODY) patients in south India. *BMC Med Genet.* **19**, 22 (2018).
34. Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
35. Moayyeri, A., Hammond, C. J., Hart, D. J. & Spector, T. D. The UK Adult Twin Registry (TwinsUK Resource). *Twin Res. Hum. Genet.* **16**, 144–149 (2013).
36. Roden, D. M. & George, A. L. Jr. The genetic basis of variability in drug responses. *Nat. Rev. Drug Discov.* **1**, 37–44 (2002).
37. Ashley, E. A. et al. Clinical assessment incorporating a personal genome. *Lancet* **375**, 1525–1535 (2010).
38. Johnson, J. A. et al. Clinical Pharmacogenetics Implementation Consortium guidelines for CYP2C9 and VKORC1 genotypes and warfarin dosing. *Clin. Pharmacol. Ther.* **90**, 625–629 (2011).
39. Karczewski, K. J., Daneshjou, R. & Altman, R. B. Chapter 7: Pharmacogenomics. *PLoS Comput. Biol.* **8**, e1002817 (2012).
40. Urban, T. J. & Goldstein, D. B. Pharmacogenetics at 50: genomic personalization comes of age. *Sci. Transl. Med.* **6**, 220ps1 (2014).
41. Johnson, J. A. et al. Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for pharmacogenetics-guided warfarin dosing: 2017 update. *Clin. Pharmacol. Ther.* **102**, 397–404 (2017).
42. Locke, A. E. et al. Exome sequencing of Finnish isolates enhances rare-variant association power. *Nature* **572**, 323–328 (2019).
43. Strauss, K. A. & Puffenberger, E. G. Genetics, medicine, and the Plain people. *Annu. Rev. Genomics Hum. Genet.* **10**, 513–536 (2009).
44. Polvi, A. et al. The Finnish disease heritage database (FinDis) update—a database for the genes mutated in the Finnish disease heritage brought to the next-generation sequencing era. *Hum. Mutat.* **34**, 1458–1466 (2013).
45. Nakatsuka, N. et al. The promise of discovering population-specific disease-associated genes in South Asia. *Nat. Genet.* **49**, 1403–1407 (2017).
46. Cox, J. J. et al. An SCN9A channelopathy causes congenital inability to experience pain. *Nature* **444**, 894–898 (2006).
47. Saleheen, D. et al. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* **544**, 235–239 (2017).
48. Steinberg, S. et al. Loss-of-function variants in ABCA7 confer risk of Alzheimer's disease. *Nat. Genet.* **47**, 445–447 (2015).
49. Chheda, H. et al. Whole-genome view of the consequences of a population bottleneck using 2926 genome sequences from Finland and United Kingdom. *Eur. J. Hum. Genet.* **25**, 477–484 (2017).
50. Lim, E. T. et al. Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet.* **10**, e1004494 (2014).
51. Nomura, A. et al. Protein-Truncating variants at the cholesterol ester transfer protein gene and risk for coronary heart disease. *Circ. Res.* **121**, 81–88 (2017).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019

## GenomeAsia100K Consortium

Jeffrey D. Wall<sup>1,47</sup>, Eric W. Stawiski<sup>2,3,44,47</sup>, Aakrosh Ratan<sup>4,47</sup>, Hie Lim Kim<sup>5,6,47</sup>, Changhoon Kim<sup>8,47</sup>, Ravi Gupta<sup>9,47</sup>, Kushal Suryamohan<sup>2</sup>, Elena S. Gusareva<sup>6</sup>, Rikky Wenang Purbojati<sup>6</sup>, Tushar Bhargale<sup>3,10</sup>, Vadim Stepanov<sup>11,12,13</sup>, Vladimir Kharkov<sup>11,12,13</sup>, Markus S. Schröder<sup>2</sup>, Vedam Ramprasad<sup>9</sup>, Jennifer Tom<sup>3</sup>, Steffen Durinck<sup>2,3</sup>, Qixin Bei<sup>2</sup>, Jiani Li<sup>2</sup>, Joseph Guillory<sup>2</sup>, Sameer Phalke<sup>9</sup>, Analabha Basu<sup>14</sup>, Jeremy Stinson<sup>2</sup>, Sandhya Nair<sup>9</sup>, Sivasankar Malaichamy<sup>9</sup>, Nidhan K. Biswas<sup>14</sup>, John C. Chambers<sup>15</sup>, Keith C. Cheng<sup>16</sup>, Joyner T. George<sup>9</sup>, Seik Soon Khor<sup>17</sup>, Jong-Il Kim<sup>18,19</sup>, Belong Cho<sup>20</sup>, Ramesh Menon<sup>9</sup>, Thirumsetti Sattibabu<sup>9</sup>, Akshi Bassi<sup>9</sup>, Manjari Deshmukh<sup>9</sup>, Anjali Verma<sup>9</sup>, Vivek Gopalan<sup>9</sup>, Jong-Yeon Shin<sup>21</sup>, Mahesh Pratapneni<sup>22</sup>, Sam Santhosh<sup>9</sup>, Katsushi Tokunaga<sup>23,24</sup>, Badrul M. Md-Zain<sup>25</sup>, Kok Gan Chan<sup>26</sup>, Madasamy Parani<sup>27</sup>, Purushothaman Natarajan<sup>27</sup>, Michael Hauser<sup>28,29</sup>, R. Rand Allingham<sup>29,46</sup>, Cecilia Santiago-Turla<sup>29</sup>, Arkasubhra Ghosh<sup>30</sup>, Santosh Gopi Krishna Gadde<sup>30</sup>, Christian Fuchsberger<sup>31,32,33</sup>, Lukas Forer<sup>33</sup>, Sebastian Schoenherr<sup>33</sup>, Herawati Sudoyo<sup>34</sup>, J. Stephen Lansing<sup>35</sup>, Jonathan Friedlaender<sup>36</sup>, George Koki<sup>37</sup>, Murray P. Cox<sup>38</sup>, Michael Hammer<sup>39</sup>, Tatiana Karafet<sup>39</sup>, Khai C. Ang<sup>16,25</sup>, Syed Q. Mehdi<sup>40,46</sup>, Venkatesan Radha<sup>41,42</sup>, Viswanathan Mohan<sup>41,42</sup>, Partha P. Majumder<sup>14,43,47\*</sup>, Somasekar Seshagiri<sup>2,45,47\*</sup>, Jeong-Sun Seo<sup>8,21,47\*</sup>, Stephan C. Schuster<sup>6,47\*</sup> & Andrew S. Peterson<sup>2,44,47\*</sup>

<sup>1</sup>Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA. <sup>2</sup>Department of Molecular Biology, Genentech, South San Francisco, CA, USA. <sup>3</sup>Department of Bioinformatics and Computational Biology, Genentech, South San Francisco, CA, USA. <sup>4</sup>Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA. <sup>5</sup>The Asian School of the Environment, Nanyang Technological University, Singapore, Singapore. <sup>6</sup>Singapore Centre for Environmental Life Sciences Engineering, Nanyang Technological University, Singapore, Singapore. <sup>7</sup>Bioinformatics Institute, Macrogen, Seoul, South Korea. <sup>8</sup>Precision Medicine Center, Seoul National University Bundang Hospital, Gyeonggi-do, South Korea. <sup>9</sup>MedGenome Labs, Bengaluru, India. <sup>10</sup>Department of Human Genetics, Genentech, South San Francisco, CA, USA. <sup>11</sup>Institute of Medical Genetics, Tomsk National Medical Research Center, Tomsk, Russian Federation. <sup>12</sup>Russian Academy of Sciences, Tomsk, Russian Federation. <sup>13</sup>Tomsk State University, Tomsk, Russian Federation. <sup>14</sup>National Institute of BioMedical Genomics, Netaji Subhas Sanatorium, Kalyani, India. <sup>15</sup>Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore. <sup>16</sup>Department of Pathology and Jake Gittlen Laboratories for Cancer Research, Penn State College of Medicine, Hershey, PA, USA.

<sup>17</sup>Department of Human Genetics, University of Tokyo, Tokyo, Japan. <sup>18</sup>Department of Biomedical Sciences, Seoul National University Graduate School, Seoul, South Korea. <sup>19</sup>Genomic Medicine Institute (GMI), Medical Research Center, Seoul National University, Seoul, South Korea. <sup>20</sup>Department of Family Medicine, Seoul National University Hospital, Seoul, South Korea. <sup>21</sup>Precision Medicine Institute, Macrogen, Gyeonggi-do, South Korea. <sup>22</sup>Emerge Ventures, Singapore, Singapore. <sup>23</sup>Genome Medical Science Project, Toyama, Japan. <sup>24</sup>National Center Biobank Network (NCBN), National Center for Global Health and Medicine (NCGM), University of Tokyo, Tokyo, Japan. <sup>25</sup>School of Environment and Natural Resource Science, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia. <sup>26</sup>Division of Genetics and Molecular Biology, Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia. <sup>27</sup>Department of Genetic Engineering, SRM Institute of Science and Technology, Kattankulathur, India. <sup>28</sup>Department of Ophthalmology, Duke University Medical Center, Durham, NC, USA. <sup>29</sup>Department of Medicine, Duke University Medical Center, Durham, NC, USA. <sup>30</sup>GROW Research Laboratory, Narayana Nethralaya Foundation, Bengaluru, India. <sup>31</sup>Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA. <sup>32</sup>Institute for Biomedicine, Eurac Research, Bolzano, Italy. <sup>33</sup>Institute of Genetic Epidemiology, Department of Genetics and Pharmacology, Medical University of Innsbruck, Innsbruck, Austria. <sup>34</sup>Genome Diversity and Diseases Laboratory, Eijkman Institute for Molecular Biology, Jakarta, Indonesia. <sup>35</sup>Complexity Institute, Nanyang Technological University, Singapore, Singapore. <sup>36</sup>Anthropology Department, Temple University, Philadelphia, PA, USA. <sup>37</sup>Papua New Guinea Institute for Medical Research, Goroka, Papua New Guinea. <sup>38</sup>School of Fundamental Sciences, Massey University, Palmerston North, New Zealand. <sup>39</sup>Division of Biotechnology, University of Arizona, Tucson, AZ, USA. <sup>40</sup>Center for Human Genetics, Sindh Institute of Urology and Transplantation, Karachi, Pakistan. <sup>41</sup>Madras Diabetes Research Foundation, Chennai, India. <sup>42</sup>Dr. Mohan's Diabetes Specialities Centre, Chennai, India. <sup>43</sup>Human Genetics Unit, Indian Statistical Institute, Kolkata, India. <sup>44</sup>Present address: Seven Rivers Genomic Medicines, A division of MedGenome, Foster City, CA, USA. <sup>45</sup>Present address: SciGenom Research Foundation, Chennai, Tamil Nadu, India. <sup>46</sup>Deceased: R. Rand Allingham, Syed Q. Mehdi. <sup>47</sup>These authors contributed equally: Jeffrey D. Wall, Eric Stawiski, Aakrosh Ratan, Hie Lim Kim, Changhoon Kim, Ravi Gupta, Partha P. Majumder, Somasekar Seshagiri, Jeong-Sun Seo, Stephan C. Schuster, Andrew S. Peterson. \*e-mail: ppm1@nibmg.ac.in; sekar@sgf.org; jeongsun@snu.ac.kr; stephan.c.schuster@gmail.com; peterson.andrew@genomeasia100k.org

## Methods

### Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to the allocation during analysis.

### Samples

We accessed publicly available high-coverage, whole-genome FASTQ files from previous studies of human genetic variation<sup>52–55</sup> and combined these with 1,267 high-coverage genomes generated as part of this project. Full details on the samples chosen for sequencing and the informed consent processes for these samples can be found in Supplementary Information 1. We restricted our analyses to genomes generated using Illumina short-read sequencing technology.

### Whole-genome sequencing

Whole-genome sequencing libraries were prepared using standard protocols (Illumina) and sequenced on Illumina HiSeq 2500/4000 or X10 machines. We obtained paired-end ( $2 \times 100$  bp or  $2 \times 150$  bp) for each sample.

### Filtering, alignment and variant calling

We aligned the Illumina short-read sequences to the GRCh37+decoy reference genome with BWA-mem<sup>56</sup> using the default parameters. Putative PCR duplicates were flagged using SAMBLASTER<sup>57</sup>. The SAM outputs were converted to BAM format, and sorted by chromosomal coordinates using Sambamba<sup>58</sup>, and all BAM files for the same samples were merged.

The sex of the samples was inferred from the coverage of the autosomes and the sex chromosomes, and confirmed from the submitted metadata with the samples. All samples that had an average coverage less than 20-fold or for which we found a difference in the inferred and reported sex were removed from further analysis. We used verifyBamID<sup>59</sup> to identify contamination using the chip-free mode and samples for which swaps or contamination was identified were removed from subsequent analyses. A contamination level of 3% was used as a cut-off, and this left us with 1,739 samples that were used for all downstream analyses.

We identified the single-nucleotide substitutions and small indels variants in the 1,739 samples using the reference model (gVCF-based) workflow for joint analysis in GATK<sup>60</sup>. Variants were called individually in each sample using the HaplotypeCaller in ‘ERC GVCF’ mode to produce a record of genotype likelihoods and annotations at each site in the genome. Multi-allelic variants are reported in the GenomeAsia browser but were not included in the analysis. A gVCF file was created for every sample, and a subsequent joint genotyping analysis of all gVCFs was done to identify the variants in the cohort. We followed the GATK-recommended best practices for variant recalibration to create a final VCF file and recalibrated the variants to select 99% of the true sites from the training set for VQSR<sup>61</sup>. The VCF files were zipped using bgzip and indexed using tabix.

### Identification of first-degree relative pairs

Several of the reported analyses require filtering to remove related samples. We used KING<sup>62</sup> to identify such first-degree relative pairs. We first used vcftools<sup>63</sup> and plink<sup>64</sup> to convert the VCF file into the required input format for KING. The estimated kinship coefficient was restricted to 0.177–0.354 as described in the KING manual to identify the first-degree relative pairs, and the results were confirmed from the submitted metadata. The number of unrelated samples by country-of-origin is shown in Supplementary Table 1.1.

### Quantifying population structure and changes in population size

We restricted our attention to 7,966,132 autosomal markers (that is, SNPs) with MAF  $\geq 0.01$  and call rate  $\geq 98\%$ . In some analysis, severe linkage disequilibrium pruning was applied as follows: sliding windows of

size 50 (that is, the number of markers used for linkage disequilibrium testing at a time) and window increments of 5 markers; for any pair of SNPs in a window, the first marker of the pair was discarded if  $r^2 > 0.2$ . After linkage disequilibrium pruning, 1,089,227 SNPs were retained for analysis. All data-filtering procedures were conducted in PLINK v.1.9<sup>64</sup>.

Analyses of population structure was performed using the quality-control-positive linkage-disequilibrium-pruned set of 1,089,227 autosomal SNPs. Principal component analysis (PCA)<sup>18</sup> was conducted across all available populations in EIGENSTRAT v.6.1.4. Results were visualized in Tableau v.9.3. We applied unsupervised hierarchical clustering of individuals using the maximum likelihood method implemented in ADMIXTURE v.1.3.0<sup>20</sup> using default input parameters. The ‘-cv’ flag was adopted to perform the cross-validation procedure and to calculate the optimal  $k$  value.

We used MSMC<sup>5</sup> to estimate changes in population size and split times. This analysis used two different phased genome datasets (using Shapeit v.2<sup>65</sup> and Eagle<sup>266</sup>). The details for the phasing are described in Supplementary Information 4. Chromosome 6 was excluded from the analysis owing to possible phasing errors in the HLA region. We used four haplotypes (two individual genomes) for estimating changes in population size in a population and eight haplotypes (two genomes from each of a pair of populations) for the estimation of population split times. We assumed a mutation rate of  $\mu = 1.25 \times 10^{-8}$  per site per generation and an average generation time of 29 years, as in previous studies<sup>8,19</sup>.

### Comparison with 1000 Genomes Project genotype calls

We filtered the variant calls to include only biallelic SNPs with  $<10\%$  missing genotype calls that were within the 1000 Genomes Project strict mask (available at [ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible\\_genome\\_masks/20141020\\_strict\\_mask.whole\\_genome.bed](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/20141020_strict_mask.whole_genome.bed)). Then, for each of the 119 overlapping samples considered individually, we calculated variant discordance rates for those filtered SNPs that (1) had a genotype call in both the 1000 Genomes Project data and the GASp data; and (2) had a ‘variant’ call (that is, a non-homozygous reference genotype call) in at least one of the datasets. These discordance rates were then stratified by the estimated MAF in the GASp dataset.

### Patterns of allele sharing

We used a parsimony-based analysis of allele sharing<sup>55</sup> that focused on SNPs that were not present in sub-Saharan Africans or in archaic humans (further details are provided in Supplementary Information 8).

### Archaic admixture

We used a method similar to the ‘enhanced’  $D$ -statistic approach<sup>8,67</sup> to estimate levels of Neanderthal and Denisovan ancestry in each non-African sample. The estimates were calibrated assuming 0% Denisovan ancestry in the British population, 4% Denisovan ancestry in the Papuan population and 2% Neanderthal ancestry in the British population (full details are provided in Supplementary Information 9).

### Determination of high-quality variants for medically related analyses

High-quality variants were defined as variants that (1) had a read-depth  $\geq 5$  and genotype-quality  $\geq 20$ ; (2) were contained in the high-confidence regions as described by Genome in a Bottle ([ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878\\_HG001/NISTv3.3.2/GRCh37/supplementaryFiles/HG001\\_GRCh37\\_GIAB\\_highconf\\_CG-IllFB-IIIgATKHC-Ion-10X-SOLID\\_CHROM1-X\\_v.3.3.2\\_highconf.bed](ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv3.3.2/GRCh37/supplementaryFiles/HG001_GRCh37_GIAB_highconf_CG-IllFB-IIIgATKHC-Ion-10X-SOLID_CHROM1-X_v.3.3.2_highconf.bed)) and (3) passed the gnomAD\_Filter. Variant annotation was carried out using SnpEff<sup>68</sup> (v.4.1).

### IBD scores

Groups with at least two samples were considered for analysis. We restricted our analysis to genomic regions with high-confidence calls

# Article

and removed related samples based on reported relationship, kinship, PCA and IBD analyses. The scores given in the figure are relative ratios compared to that of the Finnish group.

## PTVs

PTVs are defined as high-quality variants that were annotated as having a strong impact on the protein (such as frameshifts, essential splice sites or premature stop codons). We restricted calls to high-confidence regions determined by Genome in a Bottle as described above and filtered for high-confidence PTVs using the LOFTEE program<sup>69</sup>. We used a similar strategy for additional filtering of variants as proposed previously<sup>47</sup> and flagged variants with  $\leq 7$  reads covering the variant site;  $\leq 80\%$  of reads had the variant, were not in the bottom 1 percentile of phyloP or gerpRS<sup>65</sup> scores and for which the affected transcripts made up less than 50% of all expression as specified by GTEx.

## Enriched medically relevant variants

We compared variant allele counts for Asian and Oceania samples from the GenomeAsia cohort to allele counts present in non-Asian gnomAD samples (European (non-Finnish), European (Finnish), Latino, African or other) for variants found in a set of 124 medically relevant genes. The genes used were 115 genes used for prenatal screening<sup>70</sup> as well as the cancer-associated genes *BRCA1*, *BRCA2*, *TP53*, *MEN1*, *MLH1*, *MSH2*, *MSH6*, *PMS1* and *PMS2A*. A Fisher's exact test was used to calculate variations that were significantly overrepresented in the GenomeAsia subsamples and corrected for multiple testing using the Bonferroni method. We further accessed variants for these genes that had not previously been reported. All variants were further filtered as being damaging as determined by having a high impact on the protein (stop codon, essential splice site or frameshift mutation) or were predicted to be damaging by the Polyphen2 program. A cumulative comparison of allele counts for all over-represented and novel variants was performed and compared to non-Asian gnomAD to calculate a *P* value, odds ratio and relative difference in cumulative allele frequency (GenomeAsia cumulative allele frequency minus gnomAD non-Asian allele frequency). Reported *P* values were corrected for multiple testing using the Bonferroni method.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

For each variant, summary data for genotype quality, allele depth and population-specific allele counts were calculated before removing all genotype data. This dataset is available without requirement for login or other form of restriction for browsing or for download at <https://browser.genomeasia100k.org>. Individual level VCF data files representing the 1,180 newly sequenced genomes from individuals of 74 population groups are freely available to any qualified investigator without restriction. Chinese samples sequenced were from Corriell cell lines and are not subject to Chinese government regulation. The data are also available from the European Genome Archive (EGA) under accession number EGAS00001002921. The procedure for accessing individual level data are as follows: access forms can be obtained from the GenomeAsia website (<https://browser.genomeasia100k.org>), and once filled out and sent to [dataaccess@genomeasia100k.org](mailto:dataaccess@genomeasia100k.org) the request will undergo administrative review and instructions for downloading the data will be returned to the requestor. Access to individual level data from Malaysian samples are subject to additional restrictions. The complete dataset of sequences of unrelated individuals (1,667 samples) has been phased and can be used for imputation through the Michigan Imputation Server (<https://imputationserver.sph.umich.edu/index.html>). The goal of the GenomeAsia100K consortium is to

facilitate and accelerate genetic studies in Asian populations by coordinating sequencing efforts among its members. To achieve this goal, we are committed to continuing to make data publicly available and accessible. As data are contributed to the consortium by individual members, it will be made immediately available in summary form or as imputation reference panels where appropriate. Data will be made available in individual form wherever possible and not limited by the bounds of informed consent, national privacy laws and regulations, or other external restrictions that may apply.

52. Wong, L. P. et al. Deep whole-genome sequencing of 100 southeast Asian Malays. *Am. J. Hum. Genet.* **92**, 52–66 (2013).
53. Wong, L. P. et al. Insights into the genetic structure and diversity of 38 South Asian Indians from deep whole-genome sequencing. *PLoS Genet.* **10**, e1004377 (2014).
54. Vernot, B. et al. Excavating Neanderthal and Denisovan DNA from the genomes of Melanesian individuals. *Science* **352**, 235–239 (2016).
55. Wall, J. D. Inferring human demographic histories of non-African populations from patterns of allele sharing. *Am. J. Hum. Genet.* **100**, 766–772 (2017).
56. Aaboud, M. et al. Combination of the searches for pair-produced vectorlike partners of the third-generation quarks at  $\sqrt{s} = 13$  TeV with the ATLAS detector. *Phys. Rev. Lett.* **121**, 211801 (2018).
57. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).
58. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
59. Jun, G. et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91**, 839–848 (2012).
60. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
61. Van der Auwera, G. A. et al. From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).
62. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
63. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
64. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
65. Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
66. Loh, P. R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
67. Prüfer, K. et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
68. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain *w1118; iso-2; iso-3*. *Fly* **6**, 80–92 (2012).
69. MacArthur, D. G. et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
70. Haque, I. S. et al. Modeled fetal risk of genetic diseases identified by expanded carrier screening. *J. Am. Med. Assoc.* **316**, 734–742 (2016).

**Acknowledgements** We thank the many individuals from all across Asia who gave blood samples for scientific research and the many individuals who supported the sample collection. The computational work for this article was partially performed on resources of the National Supercomputing Centre, Singapore (<https://www.nssc.sg>).

**Author contributions** J.D.W., E.W.S., A.R., A. Basu, K.C.C., M. Pratapneni, S. Santhosh, H.S., J.S.L., P.P.M., J.-S.S., S.C.S., S. Seshagiri and A.S.P. designed the study. S. Seshagiri, A.S.P., S.C.S., V. Ramprasada, J.G.J.S. and J.T.G. produced the sequencing data. S. Seshagiri, S.P., J.D.W., E.W.S., R.W.P. and A.R. carried out the data processing and quality control. P.P.M., K.C.C., J.S.L., J.-S.S., S.N., S.M., J.T.G., S.S.K., S.G.K.G., K.G.C., J.-I.K., C.K., B.C., B.M.M.-Z., J.-Y.S., K.T., M. Parani, P.N., C.S.-T., M. Hauser, R.R.A., A.G., M.P.C., J.F., M. Hammer, T.K., K.C.A., S.Q.M., V.M., V. Radha and G.K. coordinated, collected and/or provided samples. C.F., L.F. and S. Schoenherr generated the imputation server. P.P.M., S. Seshagiri, J.D.W., E.W.S., A.R., A.S.P., H.L.K., R.G., K.S., E.S.G., T.B., V.K., V.S., M.S.S., J.T., S.D., Q.B., J.L., N.K.B., R.M., T.S., A.V., V.G., A. Bassi, A. Basu, C.K. and M.D. carried out analyses. J.D.W., S. Seshagiri, E.W.S. and A.S.P. wrote the paper.

**Competing interests** A.S.P., E.W.S., S. Seshagiri, T.B., J.T.G., J.T., J. Stinson, Q.B., M.S.S., S.D. and K.S. were employees of Genentech at the time this work was carried out. S. Santhosh, A.V., M. Pratapneni, V. Ramprasada, S.P., R.M., R.G., S.N., S.M., T.S., V.G., J.T.G., M.D. and S.P. are employees of and/or have equity in MedGenome. C.K., J.-S.S. and J.-Y.S. are employees of Macrogen.

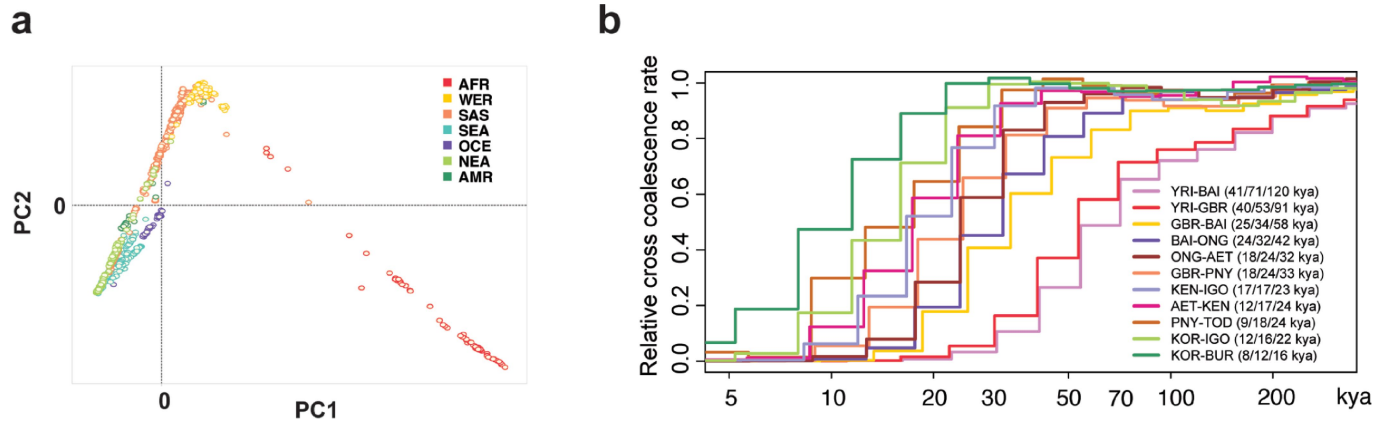
## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1793-z>.

**Correspondence and requests for materials** should be addressed to S. Seshagiri, J.-S.S., S. Schuster and A.S.P.

**Peer review information** Nature thanks Rasmus Nielsen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

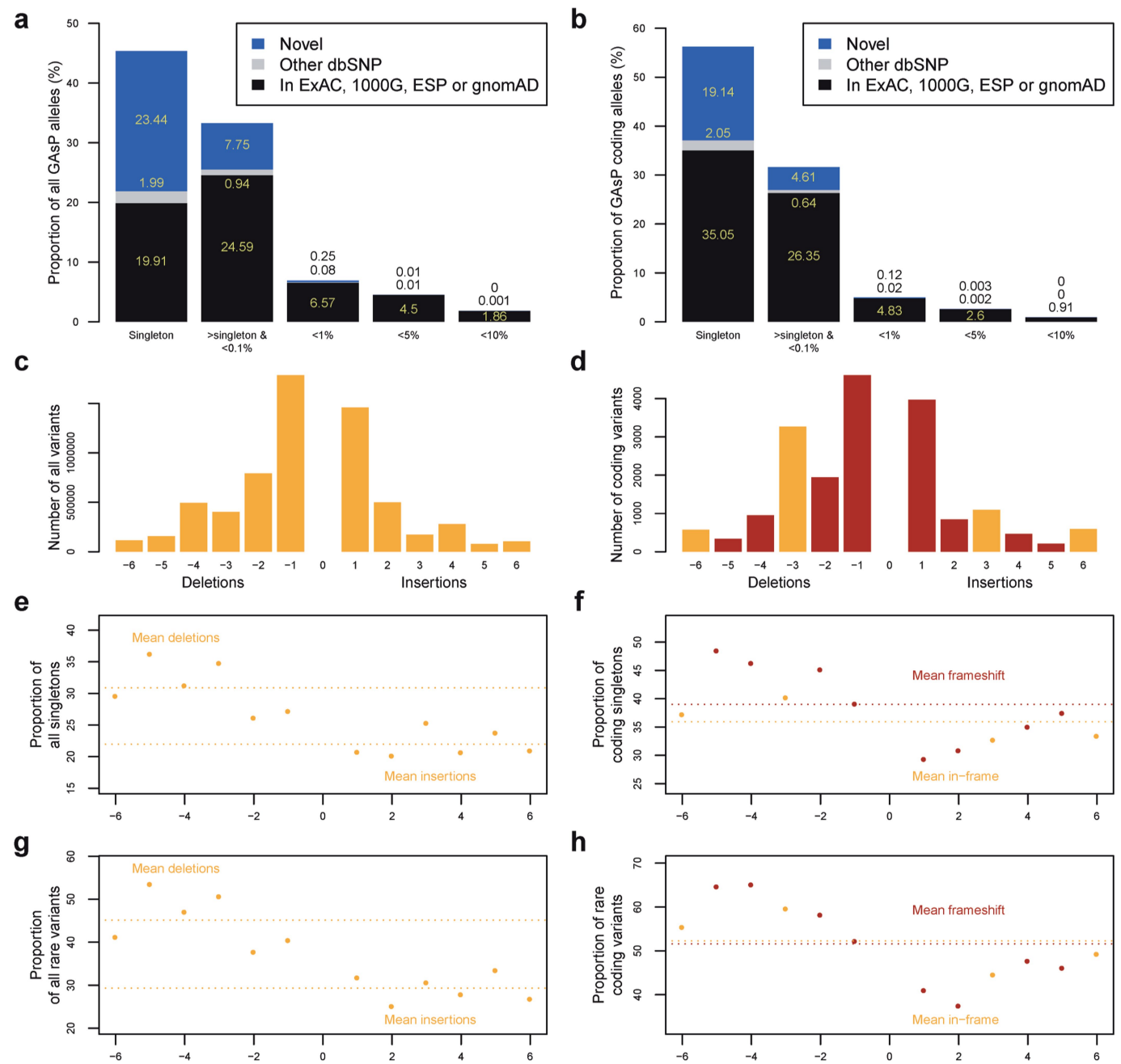
**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



**Extended Data Fig. 1 | Diversity and divergence times of GASp samples.**  
**a**, PCA plot of study samples. Africa (AFR),  $n = 102$ ; West Eurasia (WER),  $n = 111$ ; South Asia (SAS),  $n = 642$ ; Southeast Asia (SEA),  $n = 162$ ; Oceania (OCE),  $n = 68$ ; Northeast Asia (NEA),  $n = 346$ ; Americas (AMR),  $n = 26$ . The samples included in each of these geographically defined groups are described in Supplementary

Table 1a. **b**, MSMC cross-coalescence rates showing divergence time estimates between different groups. The point estimate of the date was given at which 25%, 50% and 75% of lineages in the pair of populations have coalesced into a common ancestral population.





**Extended Data Fig. 2 | Characteristics of GAsP SNPs and indels.**  
**a, b**, Comparison of all GAsP variants (**a**) or coding variants (**b**) with gnomAD, ExAC, 1000 Genomes, ESP and dbSNP data as a function of the MAF within the

GAsP dataset. **c, d**, The number and lengths of small indels in the genome (**c**) or coding regions (**d**). **e–h**, Proportion of non-coding (**e, g**) or coding (**f, h**) indels that were singletons (**e, f**) or rare (allele frequency of <0.1%; **g, h**).

# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☐ ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☐ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

## Software and code

Policy information about [availability of computer code](#)

Data collection

no software was used

Data analysis

BWA version 0.7.13 (<https://github.com/lh3/bwa>);  
 SAMBLASTER version 0.1.22 (<https://github.com/GregoryFaust/samblaster>) Sambamba version 0.6.1 (<https://github.com/lomereiter/sambamba>) BAMreport version 0.0.2; (<https://github.com/aakrosh/BAMreport>) verifyBamID version 1.1.3 (<http://genome.sph.umich.edu/wiki/VerifyBamID>); GATK version 3.5 (<https://software.broadinstitute.org/gatk/>);  
 vcfnano version 0.1.0-dev (<https://github.com/brentp/vcfnano>);  
 htlib version 1.3.1-64-g74bcfd7 (<https://github.com/samtools/htlib>); vcftools version 0.1.14 (<https://vcftools.github.io/index.html>);  
 plink version 1.90b3.40 (<http://zzz.bwh.harvard.edu/plink/>); king version 1.4 (<http://people.virginia.edu/~wc9c/KING/>); rtg-tools version 3.7 (<https://github.com/RealTimeGenomics/rtg-tools>);  
 Shapeit v2 (Delaneau et al, 2012);  
 ex- tractPIRs (Delaneau et al, 2013);  
 Eagle2 algorithm (Loh et al. 2016), version 2.3;  
 generate\_multihetsep.py, downloaded from <https://github.com/stschiff/msmc-tools>;  
 Admixture v.1.3.0 (Alexander et al, 2009);  
 EIGENSTRAT v.6.1.4 (Price et al, 2006);  
 Selscan v. 1.1.0 (Szpiech and Hernandez 2014);  
 BEAST v.1.8.4 (Drummond et al. 2012);  
 PLINK v1.9

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

For each variant, summary data for genotype quality, allele depth and population specific allele counts were calculated before removing all genotype data. This data set is available without requirement for login or other form of restriction for browsing or for download at (<https://browser.genomeasia100k.org>). Individual level VCF data files representing 1,180 newly sequenced genomes from individuals in 74 population groups are freely available to any qualified investigator without restriction. Chinese samples sequenced were from Coriell cell lines and are not subject to the Chinese regulation. The data are available from the European Genome Archive (EGA) under accession number EGAS00001002921.

The procedure for accessing individual level data is as follows:

Access forms obtained from the GenomeAsia website (<https://browser.genomeasia100k.org>), once filled out and returned to [dataaccess@genomeasia100k.org](mailto:dataaccess@genomeasia100k.org) will undergo administrative review and instructions for download will be returned to the requestor. Access to individual level data from Malaysian samples are subject to additional restrictions.

The complete data set of sequences of unrelated individuals (1,667 samples) has been phased and can be used for imputation through the Michigan Imputation Server (<https://imputationserver.sph.umich.edu/index.html>)

The goal of the GenomeAsia100K consortium is to facilitate and accelerate genetic studies in Asian populations by coordinating sequencing efforts amongst its members. To achieve this goal we are committed to continuing to make data publicly available and accessible. As data is contributed to the consortium by individual members it will be made immediately available in summary form or as imputation reference panels where appropriate. Data will be made available in individual form wherever possible and not limited by the bounds of informed consent, national privacy laws and regulations, or other external restrictions that may apply.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. and the investigators were not blinded to the allocation during analysis.
Data exclusions	data was not excluded unless it failed essential QC metrics
Replication	results were not externally replicated
Randomization	The experiments were not randomized.
Blinding	Investigators were not blinded to the allocation during analysis.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

# Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	200 populations groups were included in our study and study participants included equal numbers of both genders
Recruitment	participants were recruited based on self and external identification as member of a specific population groups
Ethics oversight	Nanyang Technological University institutional review board (IRB- 2014-12-011)

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Longitudinal molecular trajectories of diffuse glioma in adults

<https://doi.org/10.1038/s41586-019-1775-1>

Received: 8 February 2019

Accepted: 1 October 2019

Published online: 20 November 2019

Floris P. Barthel<sup>1,2,72</sup>, Kevin C. Johnson<sup>1,72</sup>, Frederick S. Varn<sup>1</sup>, Anzhela D. Moskalik<sup>1</sup>, Georgette Tanner<sup>3</sup>, Emre Kocakavuk<sup>1,4,5</sup>, Kevin J. Anderson<sup>1</sup>, Olajide Abiola<sup>1</sup>, Kenneth Aldape<sup>6</sup>, Kristin D. Alfaro<sup>7</sup>, Donat Alpar<sup>8,9</sup>, Samirkumar B. Amin<sup>1</sup>, David M. Ashley<sup>10</sup>, Pratiti Bhandopadhyay<sup>11,12</sup>, Jill S. Barnholtz-Sloan<sup>13</sup>, Rameen Beroukhi<sup>12,14</sup>, Christoph Bock<sup>8,15</sup>, Priscilla K. Brastianos<sup>16</sup>, Daniel J. Brat<sup>17</sup>, Andrew R. Brodbelt<sup>18</sup>, Alexander F. Bruns<sup>3</sup>, Ketan R. Bulsara<sup>19</sup>, Aruna Chakrabarty<sup>20</sup>, Arnab Chakravarti<sup>21</sup>, Jeffrey H. Chuang<sup>1,22</sup>, Elizabeth B. Claus<sup>23,24</sup>, Elizabeth J. Cochran<sup>25</sup>, Jennifer Connelly<sup>26</sup>, Joseph F. Costello<sup>27</sup>, Gaetano Finocchiaro<sup>28</sup>, Michael N. Fletcher<sup>29</sup>, Pim J. French<sup>30</sup>, Hui K. Gan<sup>31,32</sup>, Mark R. Gilbert<sup>33</sup>, Peter V. Gould<sup>34</sup>, Matthew R. Grimmer<sup>27</sup>, Antonio Iavarone<sup>35,36,37</sup>, Azzam Ismail<sup>20</sup>, Michael D. Jenkinson<sup>18</sup>, Mustafa Khasraw<sup>38</sup>, Hoon Kim<sup>1</sup>, Mathilde C. M. Kouwenhoven<sup>39</sup>, Peter S. LaViolette<sup>40</sup>, Meihong Li<sup>1</sup>, Peter Lichter<sup>29</sup>, Keith L. Ligon<sup>12,41</sup>, Allison K. Lowman<sup>40</sup>, Tathiane M. Malta<sup>42</sup>, Tali Mazor<sup>27</sup>, Kerrie L. McDonald<sup>43</sup>, Annette M. Molinaro<sup>27</sup>, Do-Hyun Nam<sup>44,45</sup>, Naema Nayyar<sup>16</sup>, Ho Keung Ng<sup>46</sup>, Chew Yee Ngan<sup>1</sup>, Simone P. Niclou<sup>47</sup>, Johanna M. Niers<sup>39</sup>, Houtan Noshmehr<sup>42</sup>, Javad Noorbakhsh<sup>1</sup>, D. Ryan Ormond<sup>48</sup>, Chul-Kee Park<sup>49</sup>, Laila M. Poisson<sup>50</sup>, Raul Rabadan<sup>51,52</sup>, Bernhard Radlwimmer<sup>29</sup>, Ganesh Rao<sup>53</sup>, Guido Reifenberger<sup>54</sup>, Jason K. Sa<sup>45</sup>, Michael Schuster<sup>8</sup>, Brian L. Shaw<sup>16</sup>, Susan C. Short<sup>3</sup>, Peter A. Silveis Smitt<sup>30</sup>, Andrew E. Sloan<sup>55,56,57</sup>, Marion Smits<sup>58</sup>, Hiromichi Suzuki<sup>59</sup>, Ghazaleh Tabatabai<sup>60</sup>, Erwin G. Van Meir<sup>61</sup>, Colin Watts<sup>62</sup>, Michael Weller<sup>63</sup>, Pieter Wesseling<sup>2,64</sup>, Bart A. Westerman<sup>65</sup>, Georg Widhalm<sup>66</sup>, Adelheid Woehrer<sup>67</sup>, W. K. Alfred Yung<sup>7</sup>, Gelareh Zadeh<sup>68</sup>, Jason T. Huse<sup>69,70</sup>, John F. De Groot<sup>7</sup>, Lucy F. Stead<sup>3</sup>, Roel G. W. Verhaak<sup>1\*</sup> & The GLASS Consortium<sup>71</sup>

The evolutionary processes that drive universal therapeutic resistance in adult patients with diffuse glioma remain unclear<sup>1,2</sup>. Here we analysed temporally separated DNA-sequencing data and matched clinical annotation from 222 adult patients with glioma. By analysing mutations and copy numbers across the three major subtypes of diffuse glioma, we found that driver genes detected at the initial stage of disease were retained at recurrence, whereas there was little evidence of recurrence-specific gene alterations. Treatment with alkylating agents resulted in a hypermutator phenotype at different rates across the glioma subtypes, and hypermutation was not associated with differences in overall survival. Acquired aneuploidy was frequently detected in recurrent gliomas and was characterized by IDH mutation but without co-deletion of chromosome arms 1p/19q, and further converged with acquired alterations in the cell cycle and poor outcomes. The clonal architecture of each tumour remained similar over time, but the presence of subclonal selection was associated with decreased survival. Finally, there were no differences in the levels of immunoediting between initial and recurrent gliomas. Collectively, our results suggest that the strongest selective pressures occur during early glioma development and that current therapies shape this evolution in a largely stochastic manner.

Diffuse glioma is the most common malignant brain tumour in adults and invariably relapse despite treatment with surgery, radiotherapy and chemotherapy. The molecular landscape of glioma at diagnosis has been extensively characterized<sup>3–9</sup>. Although these efforts have led to the identification of driver genes and clinically relevant subtypes<sup>10,11</sup>, how the glioma genetic landscape evolves over time and in response to therapy is unknown.

Intratumoral heterogeneity is a well-recognized characteristic of gliomas and results from selective pressures such as a limited availability of nutrients, clonal competition and treatment<sup>12–15</sup>. Tumours are thought to circumvent these growth bottlenecks by dynamic competition of subclones that result in the most favourable environment for tumour sustenance<sup>1</sup>. Recent studies have suggested that stochastic changes in clone frequency (that is, neutral evolution) and immune surveillance

A list of affiliations appears at the end of the paper.



may further contribute to the observed intratumoral heterogeneity<sup>16,17</sup>. An understanding of evolutionary dynamics at several time points is needed to develop strategies aimed at delaying or preventing the onset of tumour progression.

To investigate clonal dynamics over time and in response to therapeutic pressures, we established the Glioma Longitudinal Analysis (GLASS) Consortium. GLASS is a community-driven effort that seeks to overcome the logistical challenges in constructing adequately powered longitudinal genomic glioma datasets by pooling datasets from patients treated at institutions worldwide<sup>18</sup>. We have analysed longitudinal profiles across the three molecular glioma subtypes to identify the molecular processes active at initial and recurrent time points. These analyses identified few common features of glioma evolution across subtypes, and instead pointed towards highly variable and patient-specific trajectories of genomic alterations.

### GLASS cohort

We pooled existing and newly generated longitudinal DNA sequencing datasets from 288 patients treated at 35 hospitals (Supplementary Table 1, Extended Data Fig. 1). After applying quality filters, tumour samples from 222 patients with high-quality data in at least two time points were classified according to molecular markers into three major glioma subtypes: (1) IDH-mutant and chromosome 1p/19q co-deleted (hereafter referred to as IDH-mutant-codel;  $n = 25$ ); (2) IDH-mutant without co-deletion of chromosome 1p/19q (hereafter IDH-mutant-noncode;  $n = 63$ ); and (3) IDH-wild-type ( $n = 134$ ), in alignment with the World Health Organization (WHO) classification of tumours of the central nervous system<sup>10,11</sup>. For each patient, we selected two time-separated tumour samples, henceforth termed initial and recurrence, for further analysis.

### Mutational burdens and processes over time

We first evaluated temporal changes in mutational burden and processes to understand general patterns of glioma evolution. Mutation burdens in initial tumours were comparable with previously reported rates<sup>6,7,19</sup>. There were 2.20 mutations (single-nucleotide variants and small insertions or deletions) per megabase (Mb) for IDH-mutant-codels; 2.52 mutations per Mb for IDH-mutant-noncode; and 2.85 mutations per Mb for IDH-wild-type glioma (Fig. 1a, Extended Data Fig. 2a). Excluding DNA hypermutation cases (more than 10 mutations per Mb,  $n = 35$ ), the mutation burden increased after recurrence in 70% of the cohort (Extended Data Fig. 2a). To study changes during tumour progression, we separated mutations into three fractions: initial only, recurrence only, or shared. Notably, the mutation burdens of the private fractions, but not the shared fraction, were comparable between subtypes (Extended Data Fig. 2b). Patient age at diagnosis was significantly associated with the shared mutational burden ( $P = 1.7 \times 10^{-7}$ ), and to a lesser extent with the burden of mutations private to the initial tumour ( $P = 0.0256$ ) (Extended Data Fig. 2c). On average, a longer time to recurrence was associated with a larger increase in mutation burden ( $P = 0.0043$ , Extended Data Fig. 2d).

These fraction-specific differences in mutational burden suggested that the activity of distinct mutational processes may also be time-dependent. We therefore classified mutations in each fraction according to the Catalogue of Somatic Mutations in Cancer (COSMIC) signature database<sup>20</sup>. As expected, signature activity was closely related to subtype and fraction (Fig. 1b, Extended Data Fig. 3a). Signature 1 (ageing) was nearly always the dominant signature among shared mutations in IDH-wild-type tumours, whereas the shared fraction in IDH-mutant-noncode and IDH-mutant-codel tumours—tumour subtypes that are associated with a younger age of diagnosis—also showed a strong presence of signature 16 (unknown aetiology). Signatures 3 (double-strand break repair), 15 (mismatch repair) and 8 (unknown

aetiology) were mostly confined to the private fractions, which suggests that these processes were of lesser importance to tumour maintenance than those associated with ageing.

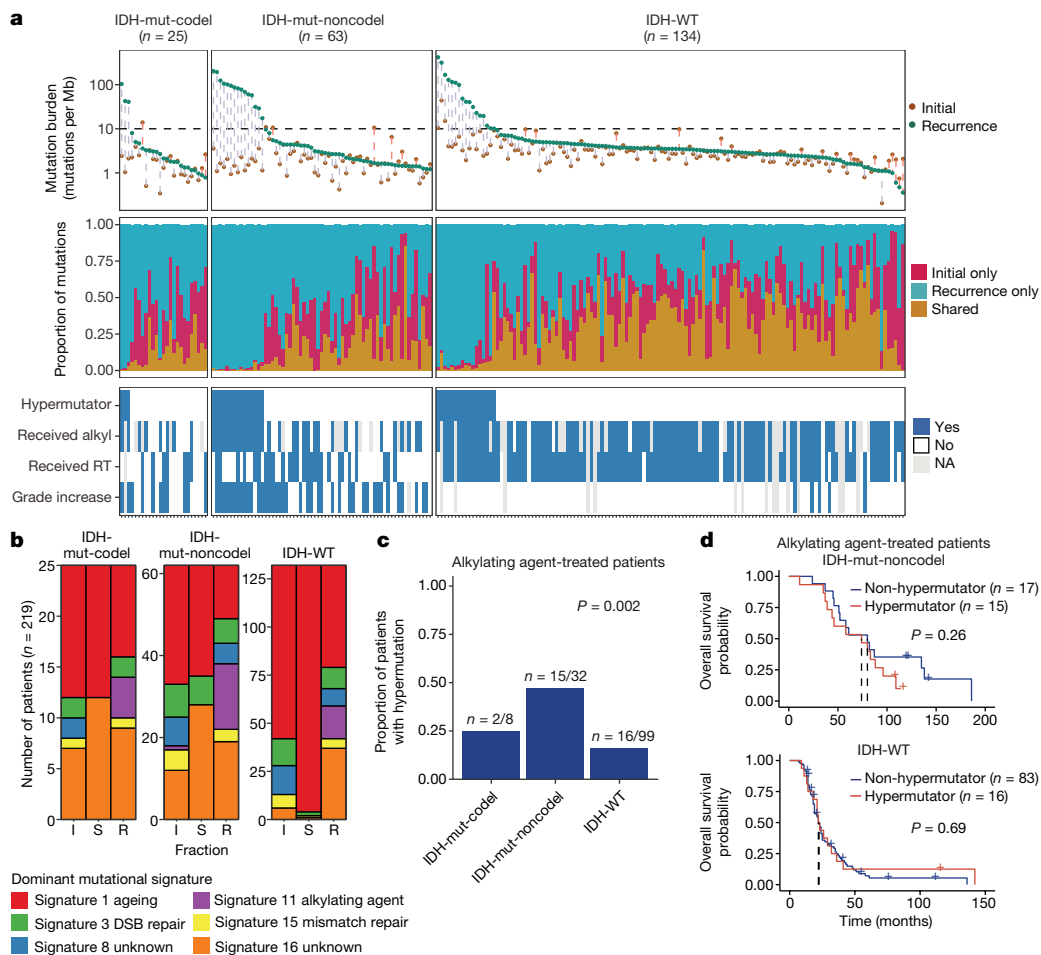
The treatment of glioma includes alkylating agents that can induce hypermutations after treatment<sup>21–23</sup>. We observed enrichment of the associated signature 11 in recurrent tumours treated with alkylating agents and with a mutational load exceeding 10 mutations per Mb (Fig. 1a, Extended Data Fig. 3b). Treatment-associated hypermutation occurred most frequently among IDH-mutant-noncode (47%), followed by IDH-mutant-codels (25%), and IDH-wild-type gliomas (16%) (Fig. 1c). The proportion of hypermutation events was significantly different between the three glioma subtypes (Fisher's exact test  $P = 2.0 \times 10^{-3}$ ), which suggests that IDH-mutant-noncode are most sensitive to developing a hypermutator phenotype<sup>24</sup>.

Treatment-induced hypermutation has been associated with disease progression<sup>23</sup>. We did not find any differences in overall survival between hypermutators and non-hypermutators treated with alkylating agents independent of age, subtype and *MGMT* methylation status (Fig. 1d, Supplementary Table 2a, b). To assess the pathogenicity of acquired mutations further, we studied their clonality<sup>25</sup>. Newly acquired clonal mutations have penetrated most of the tumour (that is, a selective sweep) between initial and recurrence and mark clonal expansion<sup>26</sup>. Conversely, acquired subclonal mutations are less prevalent, and therefore less likely to drive disease progression. Previous reports have suggested that mutations associated with alkylating agents are frequently clonal<sup>27</sup>. We found that in 48% of hypermutated tumours, most of the recurrence-only mutations were clonal, potentially reflecting cases in which a selective sweep occurred (Extended Data Fig. 4a). However, IDH-mutant-noncode hypermutators with predominantly clonal mutations did not show differences in survival compared with those containing predominantly subclonal mutations (log-rank test  $P = 0.38$ , Extended Data Fig. 4b). Alkylating agents such as temozolomide prolong the survival of adult patients with glioma<sup>28,29</sup>. Our results show that treatment-induced hypermutation is common across subtypes and does not associate with reduced overall survival, supporting the noted benefit of alkylating agent therapy.

### Selective pressures during glioma evolution

Environmental and treatment-induced pressures may drive changes in clonal architecture at recurrence. To evaluate selection over time, we clustered copy number changes and mutations on the basis of their cancer cell fraction (CCF). CCF values represent the fraction of cancer cells that contain a given alteration and reflect the relative timing of events, because alterations that are present in a subset of cancer cells probably occurred later than events present in all cancer cells (Fig. 2a). Most tumours (84%) demonstrated a mutational cluster with a CCF greater than 50% that persisted from the initial tumour to recurrence, probably reflecting the tumour trunk and containing the tumour-initiating driver mutations<sup>30</sup> (Fig. 2b, Extended Data Fig. 5a). To determine changes in clonal dominance over time, we ranked clusters within each sample by their CCF value and found similarities in clonal architecture throughout the course of disease (Kendall rank correlation,  $\tau = 0.20$ ,  $P = 3.76 \times 10^{-24}$ ; Fig. 2b, Extended Data Fig. 5b–d). These results suggested that the clonal structure at initial disease mostly persisted into recurrence.

To deepen our assessment of selective pressures, we evaluated selection in initial and recurrent tumours by determining the normalized ratio between non-synonymous and synonymous mutations (dN/dS)<sup>31</sup>. Higher ratios (above one) suggest positive selection, and ratios less than one suggest negative selection. We found evidence for positive selection at both time points despite differences between subtypes (Fig. 2c). Separating mutations into mutational fractions demonstrated that shared but not private mutations showed positive dN/dS ratios in all three glioma subtypes, which indicates that only shared



**Fig. 1 | Temporal changes in glioma mutational burden and processes.**

**a**, Each column represents a single patient ( $n = 222$ ) at two separate time points grouped by glioma subtype and ordered left-to-right by decreasing mutation frequency at recurrence. Top, mutation frequency differences between initial and recurrent tumours. Blue dotted line indicates increased mutation frequency, and the red dotted line indicates decreased mutational frequency. Middle, the proportion of total mutations shared (mustard), private to initial (magenta), or private to recurrence (blue). Bottom, clinical information including hypermutation status, therapy and grade changes. RT, radiation

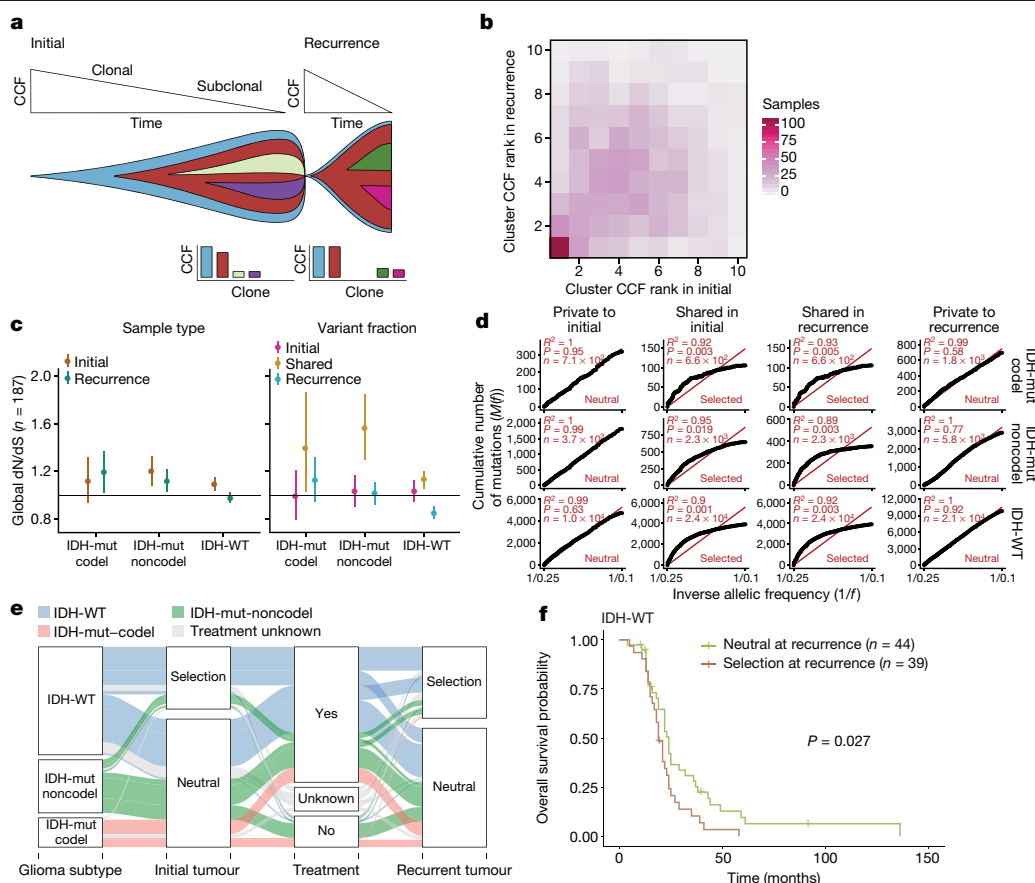
therapy. Alkyl, alkylating agent. **b**, Stacked bar plot ( $n = 219$ ) indicating the dominant mutational signature among initial, recurrent and shared mutation fractions stratified by glioma subtype. I, initial; S, shared; R, recurrence. **c**, The proportion of glioma recurrences with alkylating agent-related hypermutation, grouped by glioma subtype. Fisher's exact test was used to compare proportions between subtypes. **d**, Kaplan–Meier curve depicting overall survival in hypermutant (red) versus non-hypermutant (blue) patients treated with alkylating agent among IDH-wild-type (left,  $n = 99$ ) and IDH-mutant-noncodel (right,  $n = 32$ ) tumours.  $P$  values determined by log-rank test.

mutations (including truncal mutations) are likely to be subject to positive selection (Fig. 2c). The dN/dS ratio of initial-only mutations showed that these are neither positively nor negatively selected for, whereas recurrence-only mutations were subject to negative selection in IDH-wild-type gliomas.

To verify the reduced selective pressure in the private mutations, we used an orthogonal method to test for evidence of selection<sup>32</sup>. The method uses distributions of variant allele frequencies and estimated mutation rates to detect whether profiles significantly deviate from a model of neutral evolution (that is, as depicted by a linear relationship in Fig. 2d). In accordance with results of the dN/dS ratios, private mutations demonstrated dynamics that were consistent with neutral evolution (Fig. 2d). Shared subclonal mutations deviated from linearity and were consistent with selection both in non-hypermutators and hypermutators (Fig. 2d, Extended Data Fig. 6a, b), which provides further evidence that the strongest selective forces occur early in gliomagenesis.

Cohort-level analysis of selection masks the heterogeneity that exists in individual evolutionary trajectories. To determine the selective effects at each tumour time point, we used a Bayesian framework (SubClonalSelection algorithm) that simultaneously provides

sample-specific probabilities for both selection and neutrality while modelling sources of noise in sequencing data. The classification of a sample as 'selection' or 'neutral' is determined by whichever model has the greater probability. Classification as neutral reflects the accumulation of random mutations that are not subject to selection. Given the stringent algorithm requirements, 183 patients were included in this analysis with at least one time point, and 104 patients with both time points (16 IDH-mutant-codels, 29 IDH-mutant-noncodels, 59 IDH-wild-type; Supplementary Table 3). Neutral-to-neutral was the most common evolutionary trajectory across all three subtypes (52%), and IDH-wild-type tumours displayed the highest observed selection at any time point, with selection detected in 64% of tumours (Fisher's exact test  $P = 0.01$ ; Fig. 2e, Supplementary Table 3). IDH-wild-type gliomas with evidence for selection at recurrence had a shorter overall survival than IDH-wild-type gliomas classified as neutral at recurrence ( $P = 0.027$ ; log-rank statistic, Fig. 2f), which suggests that subclonal competition associates with more aggressive tumour behaviour. To address the limitations of smaller sample sizes in the IDH-mutant subtypes, we performed a Cox proportional hazards model including age at first diagnosis, all three glioma subtypes, and mode of selection at recurrence. This analysis revealed that selection at recurrence was significantly



**Fig. 2 | Quantifying selective pressures during glioma evolution.**

**a**, Schematic depiction of CCF values during tumour evolution indicating clonality and associated relative timing. **b**, Comparison of PyClone clusters ranked by CCF in matched initial and recurrent tumours. **c**, Left, dN/dS ratio for all variants (that is, global) in initial and recurrent tumours for each subtype. Hypermutators were not included ( $n = 187$ ). Dots represent the global dN/dS ratio with associated Wald confidence intervals. Right, global dN/dS ratios for variant fractions per subtype. **d**, Cumulative distribution of subclonal mutations by their inverse variant allele frequency. Mutations were separated

by time point, variant fraction and glioma subtype. Deviation from a linear relationship, significant Kolmogorov–Smirnov  $P$  values and Pearson's  $R^2$  values below 0.98 indicate selection. **e**, Sankey plot indicating the breakdown of SubClonalSelection evolutionary modes by subtype and therapy ( $n = 104$ ). The sizes of the bands reflect sample sizes and band colours highlight the glioma subtype. Grey colouring reflects instances when treatment information was not available. **f**, Kaplan–Meier curve showing survival differences between IDH-wild-type recurrent tumours demonstrating selection ( $n = 39$ ) compared with neutrally evolving tumours ( $n = 44$ ).  $P$  value determined by log-rank test.

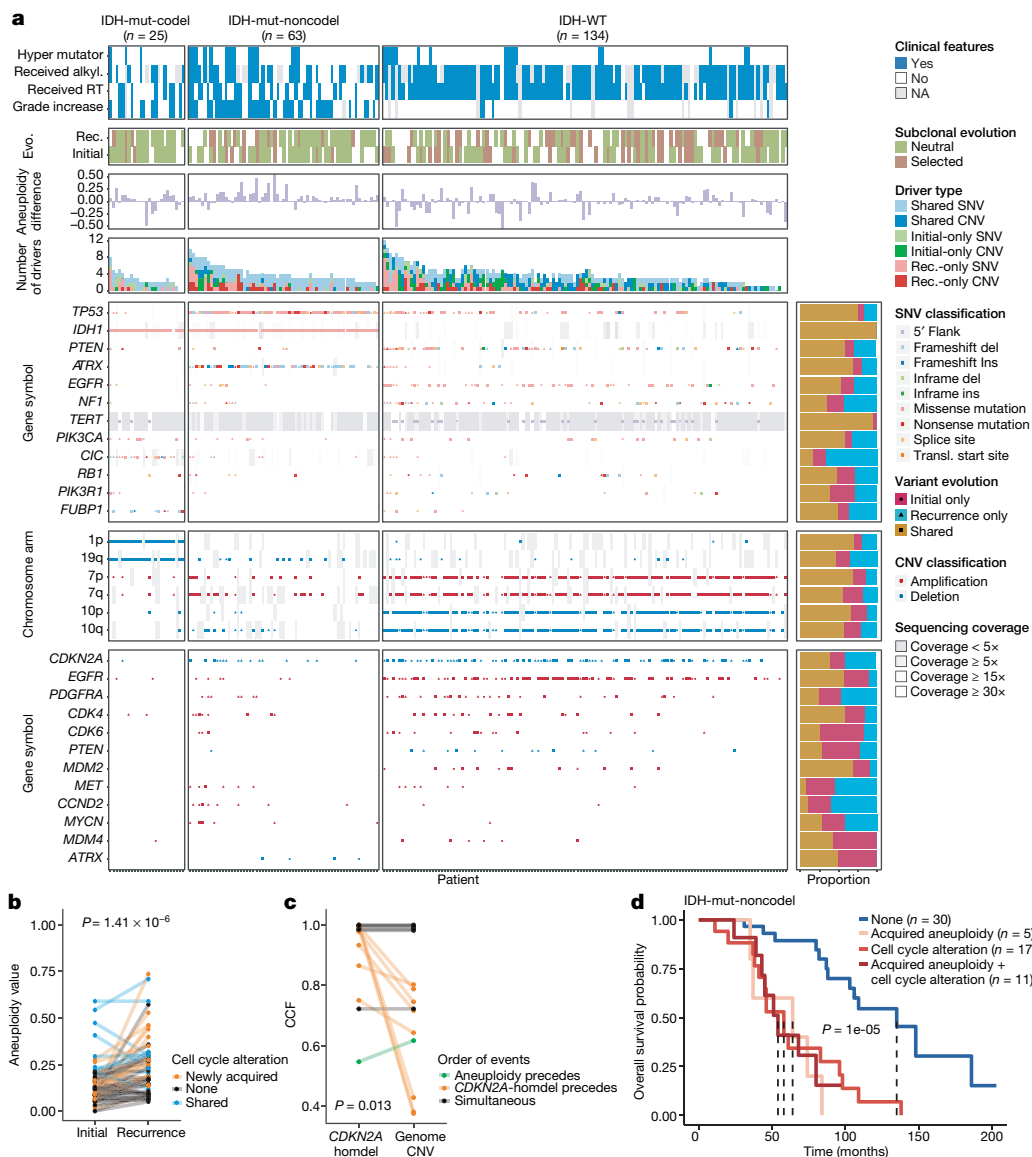
associated with shorter survival across subtypes (Hazard ratio = 1.53, 95% confidence interval 1.00–2.41,  $P = 0.048$ ; Supplementary Table 4). We next investigated whether radiation and chemotherapy imposed a selective effect, by comparing the evolutionary status at recurrence with treatment and other clinical variables. We did not observe significant associations between subclonal selection and radiation therapy or chemotherapy (Fisher's exact test  $P > 0.05$ ; Supplementary Table 5), which suggests that standard therapeutic approaches for glioma have limited effect on the subclonal tumour architecture. Although high-depth sequencing datasets may be required to detect subtle selective effects<sup>26</sup>, our analyses raise the possibility that the survival benefit derived from standard chemoradiation results from the elimination of tumour cells in which treatment sensitivity of individual cells is not determined by genetic factors.

## Driver alteration frequencies across time

We evaluated how stability, acquisition and the loss of mutation and copy number drivers<sup>6</sup> over time affect glioma evolution. We used the dN/dS ratio to nominate 12 candidate mutation driver genes at both time points ( $Q < 0.05$ , Fig. 3a, Extended Data Fig. 7a) and determined significant alterations in copy number that recapitulated previously identified drivers (Extended Data Fig. 7b). Mutations in *IDH1* and

co-occurring loss of the 1p/19q chromosome arms have been suggested as glioma-initiating events<sup>1</sup>, which was corroborated by the observation that these events were not lost or acquired during the surgical interval (Fig. 3a, Extended Data Fig. 8a). Similarly, we observed that mutations in the *TERT* promoter were almost always shared in the IDH-mutant-codel and IDH-wild-type samples, although many samples lacked sufficient coverage in this GC-rich region. Chromosome 7 gains and chromosome 10 losses were present in a large majority of IDH-wild-type initial tumours and persisted into recurrence.

Shifts in the fraction of cancer cells containing an event may also indicate a time dependency of drivers. We determined changes in cellular prevalence of shared driver events by ordering events in each sample by their CCF value (Extended Data Fig. 9). *ATRX* mutations in IDH-mutant-noncodel initial tumours demonstrated lower CCFs than *TP53* ( $P = 0.03$ ) and *IDH1* ( $P = 0.10$ ) mutations, suggesting that *IDH1* and *TP53* mutations precede *ATRX* inactivation<sup>1</sup>. There was no difference in CCF values between *IDH1* and *TP53* among initial gliomas ( $P = 0.98$ ); however, *IDH1* mutations demonstrated significantly lower CCF values than *TP53* mutations ( $P = 0.0018$ ) in recurrent gliomas. We did not observe any CCF differences among driver mutations detected in IDH-wild-type tumours at either time point. Chromosome 10 deletion CCFs were higher than chromosome 7 amplifications ( $P = 0.0036$ ), which indicates that chromosome 10 deletions arise earlier<sup>33</sup>. Similarly,



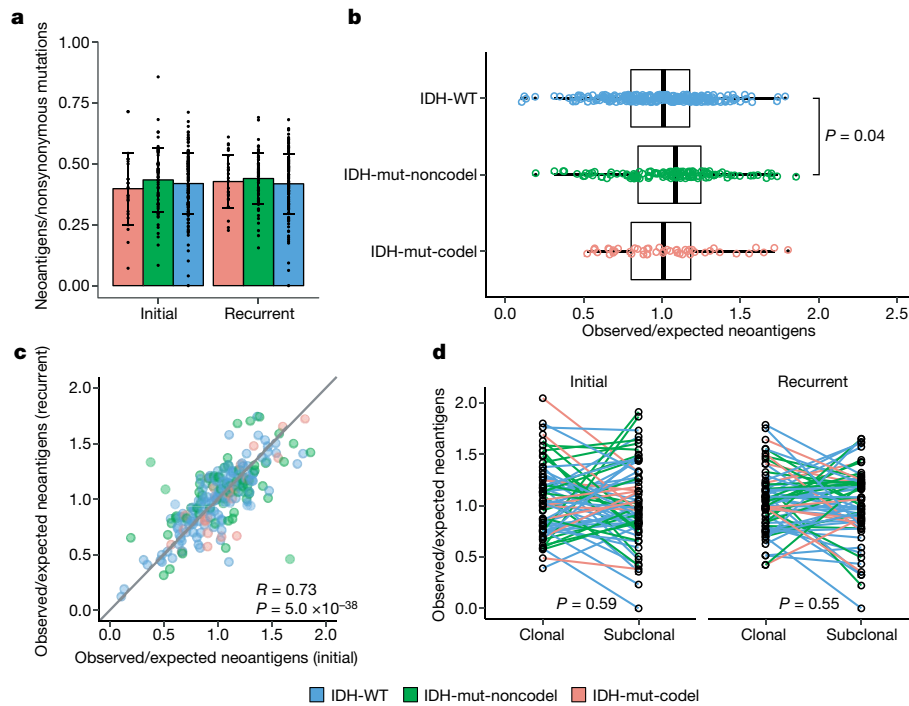
**Fig. 3 | Patterns of glioma driver frequencies over time. a**, Driver dynamics for single-nucleotide variants (SNVs) nominated by the dN/dS ratios and copy number alterations (CNVs) nominated by GISTIC ( $n = 222$ ). Each column represents a single patient at two separate time points stratified by subtype and ordered left-to-right by the number of driver alterations. The degree of aneuploidy difference (recurrence – initial) offers a summary metric for increases ( $>0$ ) or decreases ( $<0$ ) in aneuploidy at recurrence. Variants are marked and different shapes indicate whether a variant was shared or private. The variant type is depicted by its colour. Stacked bar plots accompanying each gene/arm provide cohort-level proportions for whether the alteration was

shared, lost or acquired. Rec, recurrence; evo, evolution. **b**, Aneuploidy comparison in matching initial and recurrent IDH-mutant-noncodel tumours. **c**, Within-sample CCF comparison of *CDKN2A* homozygous deletion (homdel) to genome-wide CCF as a proxy for aneuploidy. A relative higher CCF indicates temporal precedence.  $P$  value determined by Wilcoxon signed-rank test. **d**, Kaplan–Meier curve comparing survival in IDH-mutant-noncodel tumours with an alteration in the cell cycle, acquired aneuploidy, or both (shades of red) versus unaltered IDH-mutant-noncodel tumours (blue).  $P$  value determined by log-rank test.

there was no difference in CCF values between *CDKN2A* deletion and *EGFR* amplification ( $P = 0.70$ ). *EGFR* and chromosomal arm events significantly differed (that is, 10p del versus *EGFR* amp,  $P = 0.0019$ ) but not *CDKN2A* deletion and chromosomal events (that is, 10p del versus *CDKN2A* del,  $P = 0.33$ ). The consistently high CCF values for *EGFR* amplifications could indicate that these events precede even some larger chromosomal aberrations, while not excluding the possibility that high levels of extrachromosomal *EGFR*<sup>34</sup> artificially inflate CCF.

Longitudinal changes in CCF values provide additional insights into evolutionary dynamics. For instance, the CCF value may increase when a driver event is linked to clonal expansion, or conversely, decrease when a clone is outcompeted. Most individual drivers did not demonstrate significant consistent CCF changes between the initial tumour and

recurrence (Extended Data Fig. 10a). A notable exception was the *TP53* mutation CCF that increased over time ( $P = 0.037$ ) in IDH-mutant-noncodels, but not IDH-wild-type gliomas ( $P = 0.13$ , Extended Data Fig. 10b). We did not observe any differences in *IDH1* CCF over time among IDH-mutant-noncodel tumours, possibly because the general trend of these tumours to increase in CCF is counteracted by the biological loss of relevance of mutant *IDH1* over time (Extended Data Fig. 10c). Indeed, a gross comparison of all shared mutation CCFs revealed an increase in recurrent IDH-mutant-noncodel tumours ( $P < 0.0001$ ), which may reflect increased clonality and a reduction in intratumoral heterogeneity (Extended Data Fig. 10d). By contrast, shared CCFs decreased in IDH-wild-type tumours, potentially indicating a general increase in intratumoral heterogeneity at recurrence ( $P < 0.0001$ , Extended Data



**Fig. 4 | Neoantigen selection during tumour progression.** **a**, Mean proportion of coding mutations giving rise to neoantigens (neoantigens/nonsynonymous mutations) stratified by glioma subtype and time point ( $n = 222$ ). Data are mean  $\pm$  s.d. **b**, Box plot depicting the distribution of observed-to-expected neoantigen ratios in the GLASS cohort stratified by glioma subtype.  $P$  value determined by  $t$ -test. Each box spans quartiles, with the lines representing the median ratio for each group. Whiskers represent absolute range, excluding outliers. **c**, Scatterplot depicting the association between the observed-to-expected neoantigen ratio in a patient's initial versus recurrent tumours. Each point represents a single patient tumour pair.  $R$  denotes Pearson correlation coefficient. Panels **b** and **c** only include samples from pairs with at least three

neoantigens in the initial and recurrent tumours ( $n = 131$ , 63 and 24 pairs for IDH-wild-type, IDH-mutant-noncode, and IDH-mutant-codel, respectively). **d**, Ladder plot depicting the difference in observed-to-expected neoantigen ratio between a tumour's clonal and subclonal neoantigens. Each set of points connected by a line represents one tumour. Tumours are stratified by whether they were a patient's initial or recurrent tumour. Lines are coloured by each patient's glioma subtype. Panel **d** only includes samples from pairs with at least three clonal neoantigens and at least three subclonal neoantigens in both the initial and recurrent tumours ( $n = 35$ , 20 and 9 for IDH-WT, IDH-mutant-noncode and IDH-mutant-codel, respectively).  $P$  value determined by paired two-sided  $t$ -test. Colours in each panel represent the glioma subtype.

Fig. 10d). We confirmed that IDH-mutant-noncode CCF increases and IDH-wild-type decreases were not biased by patients with high mutational burden through the classification of patient-specific shared mutation CCF change (Extended Data Fig. 10e).

We next investigated whether specific somatic alterations were acquired or lost over time. Gene-specific enrichment of many recurrence-only mutations was found in hypermutated tumours, but there was no enrichment for somatic gene alterations in non-hypermutators, which suggests that glioma recurrence is not directed by particular sets of mutations (Extended Data Fig. 8b). Within subtypes, we detected an enrichment in *CDKN2A* homozygous deletions (Fig. 3a, Extended Data Fig. 8a) in recurrent IDH-mutant-noncode, which was corroborated by additional alterations to cell cycle genes (focal gain of *CCND2*, *CDK4* and *CDK6*, and mutation or homozygous loss of *RBI*). Mutations in cell cycle checkpoint control genes are associated with genomic instability<sup>35</sup>. Therefore, we analysed aneuploidy levels by determining the proportion of the genome that had undergone aneuploidy events (Extended Data Fig. 11a, b). We observed that IDH-mutant-noncode tumours had a higher level of aneuploidy at recurrence (Wilcoxon rank sum test  $P = 1.4 \times 10^{-6}$  total aneuploidy,  $P = 8.6 \times 10^{-3}$  arm-level aneuploidy; Extended Data Fig. 11c, d) with tumours carrying acquired cell cycle gene alterations displaying the largest increases in aneuploidy ( $P = 7.6 \times 10^{-6}$ ; Wilcoxon rank sum test, Fig. 3b). We reasoned that *CDKN2A* deletions may precede aneuploidy. Homozygous *CDKN2A* deletions had significantly higher CCFs than the average somatic copy number variation CCF across the genome (as a surrogate for aneuploidy-related copy number changes), suggesting that *CDKN2A* loss occurred before aneuploidy (Fig. 3c). These alterations may hasten

disease progression as patients with either alterations in cell cycle genes or the largest increases in aneuploidy at recurrence demonstrated significantly shorter survival than patients without these alterations (log-rank test  $P < 0.0001$ , Fig. 3d). Together, the persistence of drivers over time and the paucity of consistent change indicate that therapy does not result in selection of specific sets of molecular changes.

## Immunoediting activity in glioma

We next investigated how the immune microenvironment affects evolutionary trajectories. The immune system may prune tumour cells carrying immunogenic (neo-)antigens, resulting in the selection of subclones capable of evading the immune response. Evidence of this immunoediting process has been shown in several cancer types, including glioma<sup>36–39</sup>, and suggests active immunosurveillance that may be therapeutically exploited<sup>40</sup>. We computationally predicted neoantigen-causing mutations<sup>41</sup>. As expected, the neoantigen load across the GLASS cohort was strongly correlated with exonic mutation burden (Spearman's  $\rho = 0.89$ ), with 42% of nonsynonymous exonic mutations giving rise to neoantigens on average. This fraction did not significantly differ by glioma subtype or between initial and recurrent tumours ( $P > 0.05$ , Wilcoxon rank-sum test; Fig. 4a). The most common neoantigen arose from the clonal R132H mutation in *IDH1* and was present in 22 out of 88 IDH-mutant initial and recurrent tumours. Beyond mutations in *IDH1*, no mutations gave rise to a neoantigen found in more than three tumours at a given time point (Supplementary Table 6). Across the dataset, neoantigens and non-immunogenic mutations exhibited similar changes in CCF values between initial and



recurrent tumours indicating a lack of neoantigen-specific selection processes over time (Extended Data Fig. 12a).

We then examined the extent to which immunoediting occurred by comparing the observed neoantigen rate of each sample to an expected rate that was empirically derived from our dataset. The output of this approach is a normally distributed set of ratios centred at 1. Samples with an observed-to-expected neoantigen ratio less than 1 exhibit evidence of neoantigen depletion relative to the rest of the dataset, and thus are more likely to have been immunoedited. We found that none of the three glioma subtypes contained observed-to-expected ratios that significantly differed from 1 ( $P > 0.05$ , one sample  $t$ -test), although IDH-wild-type tumours exhibited significantly lower scores than IDH-mutant-noncoders ( $t$ -test,  $P = 0.04$ ; Fig. 4b). We also did not observe an association between the observed-to-expected ratio and survival when adjusting for subtype and age (Wald test,  $P > 0.05$ ), nor was there a difference between samples with neutral evolution dynamics compared to those exhibiting evidence of subclonal selection. When comparing samples longitudinally, we found that the observed-to-expected neoantigen ratio was strongly correlated between initial and recurrent tumours of each patient (Pearson's  $R = 0.73$ ,  $P = 5 \times 10^{-38}$ ), which suggests that the neoantigen depletion level in the recurrence reflects that of the initial tumour (Fig. 4c).

Immunoediting is most likely to take place in the tumours with high cytolytic activity and low levels of immunosuppressive activity<sup>39</sup>. Hypermutators, which have high loads of neoantigens, have previously been associated with highly cytolytic microenvironments<sup>38</sup>. However, we did not observe any differences in the observed-to-expected neoantigen ratio between hypermutated recurrent tumours and their initial counterparts, nor did we observe differences between hypermutated and non-hypermutated recurrent tumours, indicating that immunoediting activity is not related to the total number of mutations in a sample (Wilcoxon rank-sum test  $P > 0.05$ ; Extended Data Fig. 12b). To more directly determine whether there were immunological factors associated with neoantigen depletion, we analysed CIBERSORT immune cell fractions from a subset of samples that had undergone expression profiling in a previous study ( $n = 84$  from 42 tumour pairs)<sup>38,42</sup>. Initial tumours with an observed-to-expected neoantigen ratio greater than 1 exhibited significantly higher levels of CD4<sup>+</sup> T cells than those with a ratio less than 1, whereas recurrent tumours with a ratio greater than 1 exhibited significantly higher levels of macrophages and neutrophils, and significantly lower levels of plasma cells relative to those with ratio less than 1 ( $P < 0.05$ , Wilcoxon rank-sum test; Extended Data Fig. 12c).

Although we did not detect many factors associated with the observed-to-expected neoantigen ratio, we did observe that the ratio was significantly associated with the total number of unique HLA loci in a patient (Spearman's  $\rho = 0.28$ ,  $P = 2 \times 10^{-9}$ ), reflecting similar findings in lung cancer<sup>43</sup>. This may bias analyses comparing the ratio across patients. To determine whether immunoediting varies over time in a patient-agnostic manner, we compared the observed-to-expected neoantigen ratio derived from the clonal mutations of a sample, which likely arose earlier in tumour evolution, to that derived from their subclonal mutations, which arose later. We did not observe a significant difference in the observed-to-expected neoantigen ratio of each patient's clonal and subclonal neoantigens, regardless of glioma subtype or whether the sample was an initial tumour or recurrence ( $P > 0.05$ , paired  $t$ -test; Fig. 4d). Together, these analyses suggest that neoantigens in glioma are not exposed to differing levels of selective pressure throughout their development.

## Discussion

We reconstructed the evolutionary trajectories of 222 patients with glioma to help to understand treatment failures and tumour progression. The longitudinal molecular profiles revealed common features such as acquired hypermutation and aneuploidy, and also highlighted the

individualistic paths of glioma evolution after treatment. Our results provide evidence that the current standard of care therapies do not frequently coerce glioma down predictable paths. Instead, an unexpected number of gliomas appeared to evolve stochastically after early driver events. We expect that continuing to profile patient tumours over time using comprehensive sequencing approaches will identify other common evolutionary paths. Our results highlight the prospects of several ongoing efforts that may inform new glioma therapies.

The observation that treatment-induced hypermutation occurred across subtypes, but did not confer a detrimental effect on patient survival, leaves the clinical importance of glioma hypermutation uncertain<sup>21–24,27</sup>. Future analyses that consider the number of therapy cycles and *MGMT* DNA methylation status will help to determine factors that predispose tumours to hypermutation and identify therapies that effectively exploit the vulnerabilities of this phenotype (for example, high mutational burden). Acquired cell cycle alterations and aneuploidy in recurrent IDH-mutant-noncoders gliomas also provide a rationale to target these more aggressive phenotypes with CDK inhibitors<sup>44</sup> or with compounds that disrupt microtubule dynamics<sup>45</sup>. Finally, our analyses revealed that immunoediting activity does not vary in glioma over time, although we did observe variation between individual patients. Further molecular and immunological data are needed to fully understand the effect that this variability has on glioma evolution and to devise therapies directed at the glioma immune response<sup>17</sup>. To this end, we found that clonal neoantigens arising from the IDH1(R132H) mutation persisted from the initial tumour into the recurrence, justifying neoantigen vaccine approaches as treatments for initial and recurrent glioma<sup>46,47</sup>.

Collectively, these findings help shape our perspective on what constitutes an optimal treatment, and what approaches would result in the greatest removal or killing of glioma cells possible. Genomic characterization efforts such as The Cancer Genome Atlas (TCGA) have greatly increased our understanding of glioma biology but were limited to a single snapshot in evolutionary time. The GLASS resource provides a framework to study the patterns of glioma evolution and treatment response.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1775-1>.

1. Barthel, F. P., Wesseling, P. & Verhaak, R. G. W. Reconstructing the molecular life history of gliomas. *Acta Neuropathol.* **135**, 649–670 (2018).
2. Osuka, S. & Van Meir, E. G. Overcoming therapeutic resistance in glioblastoma: the way forward. *J. Clin. Invest.* **127**, 415–426 (2017).
3. Bettgeowda, C. et al. Mutations in *CIC* and *FUBP1* contribute to human oligodendroglioma. *Science* **333**, 1453–1455 (2011).
4. Zheng, S. et al. A survey of intragenic breakpoints in glioblastoma identifies a distinct subset associated with poor survival. *Genes Dev.* **27**, 1462–1472 (2013).
5. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
6. Ceccarelli, M. et al. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell* **164**, 550–563 (2016).
7. The Cancer Genome Atlas Research Network. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N. Engl. J. Med.* **372**, 2481–2498 (2015).
8. Verhaak, R. G. et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in *PDGFRA*, *IDH1*, *EGFR*, and *NF1*. *Cancer Cell* **17**, 98–110 (2010).
9. Yan, H. et al. *IDH1* and *IDH2* mutations in gliomas. *N. Engl. J. Med.* **360**, 765–773 (2009).
10. Louis, D. N. et al. International Society of Neuropathology—Haarlem consensus guidelines for nervous system tumor classification and grading. *Brain Pathol.* **24**, 429–435 (2014).
11. Louis, D. N. et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol.* **131**, 803–820 (2016).
12. Venteicher, A. S. et al. Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science* **355**, eaai8478 (2017).
13. Patel, A. P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).

14. Snuderl, M. et al. Mosaic amplification of multiple receptor tyrosine kinase genes in glioblastoma. *Cancer Cell* **20**, 810–817 (2011).
15. Sottoriva, A. et al. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc. Natl Acad. Sci. USA* **110**, 4009–4014 (2013).
16. Williams, M. J. et al. Quantification of subclonal selection in cancer from bulk sequencing data. *Nat. Genet.* **50**, 895–903 (2018).
17. Nejo, T. et al. reduced neoantigen expression revealed by longitudinal multiomics as a possible immune evasion mechanism in glioma. *Cancer Immunol. Res.* **7**, 1148–1161 (2019).
18. The GLASS Consortium. Glioma through the looking GLASS: molecular evolution of diffuse gliomas and the Glioma Longitudinal Analysis Consortium. *Neuro-oncol.* **20**, 873–884 (2018).
19. Hu, H. et al. Mutational landscape of secondary glioblastoma guides met-targeted trial in brain tumor. *Cell* **175**, 1665–1678.e1618 (2018).
20. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
21. Wang, J. et al. Clonal evolution of glioblastoma under therapy. *Nat. Genet.* **48**, 768–776 (2016).
22. Kim, H. et al. Whole-genome and multisector exome sequencing of primary and post-treatment glioblastoma reveals patterns of tumor evolution. *Genome Res.* **25**, 316–327 (2015).
23. Johnson, B. E. et al. Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science* **343**, 189–193 (2014).
24. Hunter, C. et al. A hypermutation phenotype and somatic MSH6 mutations in recurrent human malignant gliomas after alkylator chemotherapy. *Cancer Res.* **66**, 3987–3991 (2006).
25. Jolly, C. & Van Loo, P. Timing somatic events in the evolution of cancer. *Genome Biol.* **19**, 95 (2018).
26. Turajlic, S., Sottoriva, A., Graham, T. & Swanton, C. Resolving genetic heterogeneity in cancer. *Nat. Rev. Genet.* **20**, 404–416 (2019).
27. Choi, S. et al. Temozolomide-associated hypermutation in gliomas. *Neuro-oncol.* **20**, 1300–1309 (2018).
28. Baumber, B. G. et al. Temozolomide chemotherapy versus radiotherapy in high-risk low-grade glioma (EORTC 22033-26033): a randomised, open-label, phase 3 intergroup study. *Lancet Oncol.* **17**, 1521–1532 (2016).
29. Buckner, J. C. et al. Radiation plus procarbazine, CCNU, and vincristine in low-grade glioma. *N. Engl. J. Med.* **374**, 1344–1355 (2016).
30. Yap, T. A., Gerlinger, M., Futreal, P. A., Pusztai, L. & Swanton, C. Intratumor heterogeneity: seeing the wood for the trees. *Sci. Transl. Med.* **4**, 127ps10 (2012).
31. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e1021 (2017).
32. Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nat. Genet.* **48**, 238–244 (2016).
33. Korber, V. et al. Evolutionary trajectories of IDH(WT) glioblastomas reveal a common path of early tumorigenesis instigated years ahead of initial diagnosis. *Cancer Cell* **35**, 692–704.e612 (2019).
34. deCarvalho, A. C. et al. Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma. *Nat. Genet.* **50**, 708–717 (2018).
35. Giam, M. & Rancati, G. Aneuploidy and chromosomal instability in cancer: a jackpot to chaos. *Cell Div.* **10**, 3 (2015).
36. Marty, R., Thompson, W. K., Salem, R. M., Zanetti, M. & Carter, H. Evolutionary pressure against MHC class II binding cancer mutations. *Cell* **175**, 416–428.e413 (2018).
37. McGranahan, N. et al. Allele-specific HLA loss and immune escape in lung cancer evolution. *Cell* **171**, 1259–1271.e1211 (2017).
38. Wang, Q. et al. Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment. *Cancer Cell* **32**, 42–56.e46 (2017).
39. Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G. & Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160**, 48–61 (2015).
40. Dunn, G. P., Bruce, A. T., Ikeda, H., Old, L. J. & Schreiber, R. D. Cancer immunoevasion: from immunosurveillance to tumor escape. *Nat. Immunol.* **3**, 991–998 (2002).
41. Hundal, J. et al. pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. *Genome Med.* **8**, 11 (2016).
42. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
43. Rosenthal, R. et al. Neoantigen-directed immune escape in lung cancer evolution. *Nature* **567**, 479–485 (2019).
44. Raub, T. J. et al. Brain exposure of two selective dual CDK4 and CDK6 inhibitors and the antitumor activity of CDK4 and CDK6 inhibition in combination with temozolomide in an intracranial glioblastoma xenograft. *Drug Metab. Dispos.* **43**, 1360–1371 (2015).
45. van den Bent, M. et al. Efficacy of depatuxizumab mafodotin (ABT-414) monotherapy in patients with EGFR-amplified, recurrent glioblastoma: results from a multi-center, international study. *Cancer Chemother. Pharmacol.* **80**, 1209–1217 (2017).
46. Keskin, D. B. et al. Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. *Nature* **565**, 234–239 (2019).
47. Schumacher, T. et al. A vaccine targeting mutant IDH1 induces antitumour immunity. *Nature* **512**, 324–327 (2014).
- 5Department of Neurosurgery, University Hospital Essen, Essen, Germany. 6Laboratory of Pathology, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA. 7Department of Neuro-Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. 8CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria. 91st Department of Pathology and Experimental Cancer Research, Semmelweis University, Budapest, Hungary. 10Preston Robert Tisch Brain Tumor Center at Duke, Duke University Medical Center, Durham, NC, USA. 11Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. 12Broad Institute, Cambridge, MA, USA. 13Department of Population and Quantitative Health Sciences, Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, OH, USA. 14Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. 15Department of Laboratory Medicine, Medical University of Vienna, Vienna, Austria. 16Division of Neuro-Oncology, Massachusetts General Hospital, Boston, MA, USA. 17Department of Pathology, Northwestern University Feinberg School of Medicine, Chicago, IL, USA. 18Department of Neurosurgery, University of Liverpool & Walton Centre NHS Trust, Liverpool, UK. 19Division of Neurosurgery, The University of Connecticut Health Center, Farmington, CT, USA. 20Department of Cellular and Molecular Pathology, Leeds Teaching Hospital NHS Trust, St James's University Hospital, Leeds, UK. 21Department of Radiation Oncology, The Ohio State Comprehensive Cancer Center—Arthur G. James Cancer Hospital, Columbus, OH, USA. 22Department of Genetics and Genome Sciences, UConn Health, Farmington, CT, USA. 23Yale University School of Public Health, New Haven, CT, USA. 24Department of Neurosurgery, Brigham and Women's Hospital, Boston, MA, USA. 25Department of Pathology & Laboratory Medicine, Medical College of Wisconsin, Milwaukee, WI, USA. 26Department of Neurology, Medical College of Wisconsin, Milwaukee, WI, USA. 27Department of Neurosurgery, University of California San Francisco, San Francisco, CA, USA. 28Fondazione IRCCS Istituto Neurologico Besta, Milano, Italy. 29Division of Molecular Genetics, Heidelberg Center for Personalized Oncology, German Cancer Research Consortium, German Cancer Research Center (DKFZ), Heidelberg, Germany. 30Department of Neurology, Erasmus MC – University Medical Center Rotterdam, Rotterdam, The Netherlands. 31Olivia Newton-John Cancer Research Institute, Austin Health, Melbourne, Victoria, Australia. 32La Trobe University School of Cancer Medicine, Heidelberg, Victoria, Australia. 33Neuro-Oncology Branch, National Institutes of Health, Bethesda, MD, USA. 34Anatomic Pathology Service, Hôpital de l'Enfant-Jésus, CHU de Québec-Université Laval, Québec, Québec, Canada. 35Department of Neurology, Columbia University Medical Center, New York, NY, USA. 36Department of Pathology and Cell Biology, Columbia University Medical Center, New York, NY, USA. 37Institute for Cancer Genetics, Columbia University Medical Center, New York, NY, USA. 38Cooperative Trials Group for Neuro-Oncology (COGNO) NHMRC Clinical Trials Centre, The University of Sydney, Sydney, New South Wales, Australia. 39Department of Neurology, Brain Tumor Center Amsterdam, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. 40Department of Radiology, Medical College of Wisconsin, Milwaukee, WI, USA. 41Department of Oncologic Pathology, Dana-Farber Cancer Institute, Boston, MA, USA. 42Department of Neurosurgery, Henry Ford Health System, Henry Ford Cancer Institute, Detroit, MI, USA. 43Cure Brain Cancer Biomarkers and Translational Research Group, Prince of Wales Clinical School, University of New South Wales, Sydney, New South Wales, Australia. 44Department of Neurosurgery, Sungkyunkwan University School of Medicine, Samsung Medical Center, Seoul, South Korea. 45Institute for Refractory Cancer Research, Samsung Medical Center, Seoul, South Korea. 46Department of Anatomical and Cellular Pathology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, Hong Kong. 47Department of Oncology, Luxembourg Institute of Health, Luxembourg, Luxembourg. 48Department of Neurosurgery, University of Colorado School of Medicine, Aurora, CO, USA. 49Department of Neurosurgery, Seoul National University College of Medicine, Seoul National University Hospital, Seoul, South Korea. 50Department of Public Health Sciences, Henry Ford Health System, Henry Ford Cancer Institute, Detroit, MI, USA. 51Department of Biomedical Informatics, Columbia University Medical Center, New York, NY, USA. 52Department of Systems Biology, Columbia University, New York, NY, USA. 53Department of Neurosurgery, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. 54Institute of Neuropathology, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. 55Department of Neurological Surgery, University Hospitals Cleveland Medical Center, Case Western Reserve University, Cleveland, OH, USA. 56Department of Neurosurgery, Case Western Reserve University, Cleveland, OH, USA. 57Seidman Cancer Center and Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH, USA. 58Department of Radiology & Nuclear Medicine, Erasmus MC – University Medical Center Rotterdam, Rotterdam, The Netherlands. 59The Hospital for Sick Children, Toronto, ON, Canada. 60Interdisciplinary Division of Neuro-Oncology, Hertie Institute for Clinical Brain Research, DTK Partner Site Tübingen, Eberhard Karls University Tübingen, Tübingen, Germany. 61Department of Neurosurgery, School of Medicine and Winship Cancer Institute of Emory University, Atlanta, GA, USA. 62Institute of Cancer Genome Sciences, Department of Neurosurgery, University of Birmingham, Birmingham, UK. 63Department of Neurology, University Hospital Zurich, Zurich, Switzerland. 64Princess Máxima Center for Pediatric Oncology, Utrecht, The Netherlands. 65Department of Neurosurgery, Brain Tumor Center Amsterdam, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. 66Department of Neurosurgery, Medical University of Vienna, Vienna, Austria. 67Institute of Neurology, Medical University of Vienna, Vienna, Austria. 68Division of Neurosurgery, Department of Surgery, University Health Network, Toronto, Ontario, Canada. 69Department of Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. 70Department of Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. 71A list of participants and their affiliations appears in the online version of the paper. 72These authors contributed equally: Floris P. Barthel, Kevin C. Johnson. \*e-mail: roel.verhaak@jax.org

© The Author(s), under exclusive licence to Springer Nature Limited 2019

<sup>1</sup>The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA. <sup>2</sup>Department of Pathology, Brain Tumor Center Amsterdam, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. <sup>3</sup>Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK. <sup>4</sup>DKFZ Division of Translational Neurooncology at the West German Cancer Center, German Cancer Consortium Partner Site, University Hospital Essen, Essen, Germany.

Floris P. Barthel<sup>1,2,72</sup>, Kevin C. Johnson<sup>1,72</sup>, Frederick S. Varn<sup>1</sup>, Anzhela D. Moskalik<sup>1</sup>, Georgette Tanner<sup>3</sup>, Emre Kocakavuk<sup>1,4,5</sup>, Kevin J. Anderson<sup>1</sup>, Kenneth Aldape<sup>6</sup>, Kristin D. Alfaro<sup>7</sup>, Samirkumar B. Amin<sup>1</sup>, David M. Ashley<sup>10</sup>, Pratiti Bandopadhyay<sup>11,12</sup>, Jill S. Barnholtz-Sloan<sup>13</sup>, Rameen Beroukhi<sup>12,14</sup>, Christoph Bock<sup>8,15</sup>, Priscilla K. Brastianos<sup>16</sup>, Daniel J. Brat<sup>17</sup>, Andrew R. Brodbelt<sup>18</sup>, Ketan R. Bulsara<sup>19</sup>, Aruna Chakrabarty<sup>20</sup>, Jeffrey H. Chuang<sup>1,22</sup>, Elizabeth B. Claus<sup>23,24</sup>, Elizabeth J. Cochran<sup>25</sup>, Jennifer Connelly<sup>26</sup>, Joseph F. Costello<sup>27</sup>, Gaetano Finocchiaro<sup>28</sup>, Michael N. Fletcher<sup>29</sup>, Pim J. French<sup>30</sup>, Hui K. Gan<sup>31,32</sup>, Mark R. Gilbert<sup>33</sup>, Peter V. Gould<sup>34</sup>, Antonio Iavarone<sup>35,36,37</sup>, Azzam Ismail<sup>20</sup>, Michael D. Jenkinson<sup>18</sup>, Mustafa Khasraw<sup>38</sup>, Hoon Kim<sup>1</sup>, Mathilde C. M. Kouwenhoven<sup>39</sup>, Peter S. LaViolette<sup>40</sup>, Peter Lichter<sup>29</sup>, Keith L. Ligon<sup>12,41</sup>, Allison K. Lowman<sup>40</sup>, Tathiane M. Malta<sup>42</sup>, Kerrie L. McDonald<sup>43</sup>, Annette M. Molinaro<sup>27</sup>, Do-Hyun Nam<sup>44,45</sup>, Ho Keung Ng<sup>46</sup>, Simone P. Niclou<sup>47</sup>, Johanna M. Niers<sup>39</sup>, Houtan Noushmehr<sup>42</sup>, D. Ryan Ormond<sup>48</sup>, Chul-Kee Park<sup>49</sup>, Laila M. Poisson<sup>50</sup>, Raul Rabadan<sup>51,52</sup>, Bernhard Radlwimmer<sup>29</sup>, Ganesh Rao<sup>53</sup>, Guido Reifenberger<sup>54</sup>, Jason K. Sa<sup>45</sup>, Susan C. Short<sup>3</sup>, Peter A. Sillevs Smitt<sup>30</sup>, Andrew E. Sloan<sup>55,56,57</sup>, Marion Smits<sup>58</sup>, Hiromichi Suzuki<sup>59</sup>, Ghazaleh Tabatabai<sup>60</sup>, Erwin G. Van Meir<sup>61</sup>, Colin Watts<sup>62</sup>, Michael Weller<sup>63</sup>, Pieter Wesseling<sup>2,64</sup>, Bart A. Westerman<sup>65</sup>, Adelheid Woehrer<sup>67</sup>, W. K. Alfred Yung<sup>7</sup>, Gelareh Zadeh<sup>68</sup>, Jason T. Huse<sup>69,70</sup>, John F. De Groot<sup>7</sup>, Lucy F. Stead<sup>3</sup> & Roel G. W. Verhaak<sup>1\*</sup>

## Methods

### Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

### DNA sequencing and data collection

The GLASS dataset consists of both unpublished and published sequencing data as outlined in Supplementary Table 1. Among the cohort were exomes from 436 glioma samples (200 patients), whole-genome data from 165 glioma samples (78 patients), with overlapping exome/whole-genome data on 78 glioma samples (38 patients). A matching germline sequence was available for all patients. The dataset includes 257 sets of at least two time-separated tumour samples, 17 standalone recurrences, and 19 patients with at least two geographically distinct tumour portions. More specifically, the dataset includes exome or whole-genome sequencing data on 211 primary gliomas, 234 first recurrences, 32 second recurrences, 11 third recurrences and 1 fourth recurrence (Supplementary Table 7).

Newly generated whole-genome sequencing data for the Chinese University of Hong Kong (HK), Northern Sydney Cancer Centre (NS) and MD Anderson Cancer Center (MD) cohorts were subjected to 150 base paired-end sequencing. The HK samples were sequenced using HiSeqX, whereas the NS and MD cohorts were sequenced using NovaSeq, according to Illumina's protocols. Whole-exome capture was performed using the following platforms as reported in previous publications<sup>7,21–23,48–52</sup>.

The Agilent SureSelect Human All Exon 50 Mb capture kit was used for patients SF-0001–SF-0021, and the Agilent SureSelect Human All Exon V4 capture kit was used for patients SF-0024–SF-0029 in the University of California San Francisco cohort. The Agilent SureSelect Human All Exon v4 or v5 kit was used to capture samples in the Kyoto University cohort. The Samsung Medical Center cohort reported using the Agilent SureSelect kit for patients SM-R056–SM-R071, SM-R075, SM-R076 and SM-R095–SM-R114, whereas the Illumina TruSeq Exome-capture kit was used for patient SM-R072. Exome capture was performed using the Agilent SureSelect Human All Exon 50 Mb kit in the TCGA glioblastoma (GBM) cohort and the Agilent SureSelect Human All Exon v.2.0 44 Mb kit in the TCGA low grade glioma (LGG) cohort. Columbia University cases were captured using the Agilent V3 50 Mb kit, sequencing 90 bp paired-end reads for samples R009-TP, R009-R1, R011-TP, R011-R1, R014-TP, R014-R1, R017-R1, R018-R1 and R019-R1. Mapping files of initial tumour and normal samples of patients R017–R019 were obtained from the TCGA through the CG-hub. All other samples were captured using the Agilent SureSelect XT Human All Exon v.4 Kit, 80 million paired-end reads, 150× on-target coverage. Samples in the Henry Ford Hospital cohort were multiplexed and sequenced using Illumina HiSeq 2000 by the Sequencing and Microarray Facility at an average target exome coverage of 100× using 76-bp paired-end reads. Samples in the HK cohort were subjected to 75 base paired-end sequencing for HK-0001–HK-0004, as performed using NextSeq in high output mode. In the Leeds Cohort (LU), the SureSelectXT V5 kit (PE100) was used to construct exome libraries. The Illumina TruSeq Exome capture kit was used for samples at the Medical University of Vienna – Research Center for Molecular Medicine (CeMM).

### GLASS identifiers

A GLASS barcode system was created, based on TCGA barcode design, in an effort to de-identify patient information and provide an organized framework for the different pieces of the dataset.

GLASS barcodes are composed of 24 characters. The first four characters specify the project (either GLSS or TCGA). All datasets submitted to The GLASS Consortium, published and unpublished, were given the GLSS project ID. Samples that were part of the TCGA

cohorts (TCGA-GBM and TCGA-LGG) were given a TCGA designation. The next two characters designate the centre where the samples were either acquired or sequenced (Supplementary Table 7). This is followed by the four-character centre-specific patient identification that was kept as close as possible to the patient identification provided by the collaborators to allow a simplified trace-back process. Patient data are divided by a relative sample type, such as initial tumour (TP), recurrent tumour (R1), normal tissue (NB or NM, for example), or metastatic tumour sample (M1). If there was more than one recurrence the relative number was specified following 'R'. Some patients had surgeries for which a biospecimen was unavailable. Thus, a surgical number was also provided to indicate temporal ordering (Supplementary Table 8). To include spatially separated samples the portion designation was added, which is followed by one character specifying the type of analyte, either DNA (D) or RNA (R). As there is variation in the sequencing analysis, a three-character designation represents either whole-genome sequencing (WGS) or whole-exome sequencing (WXS). The last part of the GLASS barcode is a six-character designation unique to each barcode that was randomly generated.

### Computational pipelines

All pipelines were developed using snakemake 5.2.2<sup>53</sup>. Unless otherwise stated, all tools mentioned are part of the GATK 4 suite<sup>54</sup>. All data were collected at a central location (The Jackson Laboratory) and analysed using homogenous pipelines capable of processing raw fastq files as well as re-processing previously analysed bam files.

### Alignment and pre-processing

Data pre-processing was conducted in accordance to the GATK Best Practices using GATK 4.0.10.1. In brief, aligned BAM files were separated by read group, sanitized and stripped of alignments and attributes using 'RevertSam', giving one unaligned BAM (uBAM) file per readgroup. Uniform readgroups were assigned to uBAM files using 'AddOrReplaceReadgroups'. Similarly, unaligned fastq files were assigned uniformly designated readgroup attributes and converted to uBAM format using 'FastqToSam'. uBAM files underwent quality control using 'FastQC 0.11.7'. Sequencing adapters were marked using 'MarkIlluminaAdapters'. uBAM files were finally reverted to interleaved fastq format using 'SamToFastq', aligned to the b37 genome (human\_g1k\_v37\_decoy) using 'BWA MEM 0.7.17', attributes were restored using 'MergeBamAlignment'. 'MarkDuplicates' was then used to merge aligned BAM files from multiple readgroups and to mark PCR and optical duplicates across identical sequencing libraries. Lastly, base recalibration was performed using 'BaseRecalibrator' followed by 'ApplyBQSR'. Coverage statistics were gathered using 'CollectWgsMetrics'. Alignment quality control was performed running 'ValidateSamFile' on the final BAM file and quality control results were inspected using 'MultiQC 1.6a0'<sup>55</sup>. A haplotype database for fingerprinting was generated using a modified version of the code on [https://github.com/naumanjaved/fingerprint\\_maps](https://github.com/naumanjaved/fingerprint_maps). The tool 'CrosscheckFingerprints' was used to confirm that all readgroups within a sample belong to the same individual, and that all samples from one individual match. Any mismatches were marked and excluded from further analysis.

### Variant detection

Variant detection was performed in accordance to the GATK Best practices using GATK 4.1.0.0. Germline variants were called from control samples using Mutect2 in artefact detection mode and pooled into a cohort-wide panel of normals. Somatic variants were subsequently called in individual tumour samples (single-sample mode) and in entire patients using GATK 4.1 Mutect2 in multi-sample mode. Mutect2 was given matched control samples, the aforementioned panel of normals and the gnomAD germline resource as additional controls. Cross-sample contamination was evaluated using 'GetPileupSummaries' and 'CalculateContamination' run for both tumour and matching control

# Article

samples. Read orientation artefacts were evaluated using 'CollectFIR2Counts' and 'LearnReadOrientationModel'. Somatic likelihood, read orientation, sequence context, germline and contamination filters were applied using 'FilterMutectCalls'.

## Variant post-processing

BCFTools 1.9 was used to normalize, sort and index variants<sup>56</sup>. A consensus VCF was generated from all variants in the cohort, removing any duplicate variants. The consensus VCF file was annotated using GATK 4.1 Funcotator and the v1.6.20190124s annotation data source. Allele frequencies from multi-sample Mutect2 were used to compare allele frequencies between related samples. Multi-sample Mutect2 calls and filters mutations across a patient as a whole and does not determine mutation calls in a single sample. Single-sample mutation calls were overlaid on the multi-sample calls to infer whether variants were called in individual samples. Single-sample called variants that were not present in the multi-sample callset were discarded.

## Mutational burden

Mutational burden was calculated as the number of mutations per Mb sequenced. A minimum coverage threshold of 15× was required for each base. DNA hypermutation was defined for recurrent tumours with greater than 10 mutations per Mb sequenced as these values were considered outliers (1.5 times the interquartile range above the upper quartile). Notably, there were a few initial gliomas that demonstrated a mutational frequency above 10 mutations per Mb. However, the 'hypermutation' classification was restricted to only patients with this level at recurrence since these likely reflect different evolutionary paths.

## Mutational signatures

The relative contributions of the COSMIC mutational signatures were determined from a patient's initial-only, recurrence-only, and shared mutations by solving the non-negative-least squares problem for each set of mutations using the 30 signatures from version 2 (March 2015). Six signatures were dominantly enriched in at least 3% of the fractions and we resolved the non-negative-least squares problems using the reduced six-signature model to increase accuracy and reduce noise.

## Copy number segmentation

Copy number identification was performed according to the GATK Best Practices and is outlined briefly here. The pipeline differs slightly for whole genomes and whole exomes. For whole genomes, the genome was segmented into 10kb bins using 'PreprocessIntervals'. For exomes, overlapping regions between several commonly used capture kits (Broad Human Exome b37, Nextera Rapid Capture, TruSeq Exome, SeqCap EZ Exome V3, Agilent SureSelect V4, Agilent SureSelect V7) were identified using 'bedtools multiIntersectBed'. The tool 'PreprocessIntervals' was used to apply 1-kb padding and to merge overlapping intervals. In parallel, 'SelectVariants' was used to subset the gnomAD resource of germline variants to variants with a population allele frequency greater than 5%. Next, 'CollectReadcounts' was used to count reads in the bins generated by 'PreprocessIntervals' separately for autosomes and allosomes. In parallel, 'CollectAllelicCounts' was used to count reference and alternate reads at gnomAD variant sites with a population allele frequency greater than 5%. The cohort was subsequently split into batches determined by sequencing centre and 'CreateReadCountPanelOfNormals' was used to create a panel of normal for each batch. Panel of normals were created separately for allosomes and autosomes, and allosomes were separated further by sex. To improve the panel of normals further, GC content annotation of each interval as determined by 'AnnotateIntervals' were given. Next, 'DenoiseReadCounts' was used to denoise the binned readcounts output by 'CollectReadCounts', given a panel of normal determined by

batch, chromosomes (allosomes or autosomes) and sex. Denoised copy ratios were plotted and inspected for quality concerns using 'PlotDenoisedCopyRatios'. The tool 'ModelSegments' is an implementation of a gaussian-kernel binary-segmentation algorithm and was used to merge contiguous segments and assign copy and allelic ratios. The results of this segmentation were plotted using 'PlotModelledSegments' and inspected for quality concerns.

## Copy number calling

A copy number caller loosely based on GATK 'CallCopyRatioSegments' (which in turn is based off of ReCapSeg) and GISTIC was implemented to call both arm-level and high-level copy number changes, respectively<sup>57,58</sup>.

Segments (from 'ModelSegments') with a non-log<sub>2</sub> copy ratio between 0.9 and 1.1 were determined to be neutral. These segments were then weighted by length and a weighted mean and standard deviation non-log<sub>2</sub> copy ratio (once-filtered) were determined again. Outlier segments are removed and once again a weighted mean and standard deviation non-log<sub>2</sub> copy ratio (twice-filtered) were determined. Segments with a non-log<sub>2</sub> copy ratio between 0.9 and 1.1 and segments within two standard deviations of the twice-filtered mean were determined to be neutral, and segments outside of these boundaries were determined to have a low-level amplification or deletion, depending on the direction.

The weighted mean and standard deviation of the non-log<sub>2</sub> copy ratio (once-filtered) was then determined individually for each chromosome arm. Outlier segments were removed and the weighted mean and standard deviation of the non-log<sub>2</sub> copy ratio (twice-filtered) was determined again. To determine a high-level amplification and deletion threshold, the most highly amplified and deleted chromosome arms were selected, respectively. The twice-filtered mean plus (high level amplification) or minus (high level deletion) two times the standard deviation of the selected arms were used as high-level thresholds.

Gene level copy numbers were called by intersecting the gene boundaries with the segment intervals and by calculating the weighted non-log<sub>2</sub> copy ratio for that gene. The copy number call for that gene was then determined by comparing the gene-level non-log<sub>2</sub> copy ratio to the previously determined thresholds.

## dNdScv

The dN/dS ratios were estimated using the R package dNdScv<sup>31</sup> (<https://github.com/im3sanger/dndscv>) was run using the default and recommended parameters for all mutations in initial tumour samples, recurrent tumour samples, and for each mutational fraction (unique to initial, unique to recurrent and shared). All analyses were conducted separately within the three main tumour subtypes.

## Aneuploidy calculation

The most reductive metric of aneuploidy was computed by taking the size of all non-neutral segments divided by the size of all segments. The resulting aneuploidy value indicates the proportion of the segmented genome that is non-diploid.

In parallel, an arm-level aneuploidy score modelled after a previously described method was computed<sup>59</sup>. In brief, adjacent segments with identical arm-level calls (−1, 0 or 1) were merged into a single segment with a single call. For each merged/reduced segment, the proportion of the chromosome arm it spans was calculated. Segments spanning greater than 80% of the arm length resulted in a call of −1 (loss), 0 (neutral) or +1 (gain) to the entire arm, or 'NA' if no contiguous segment spanned at least 80% of the arm's length. For each sample the number of arms with a non-neutral event was finally counted. The resulting aneuploidy score is a positive integer with a minimum value of 0 (no chromosomal arm-level events detected) and a maximum value of 39 (total number of autosomal chromosome arms excluding the short arms for chromosomes 13, 14, 15, 21 and 22).



## Estimates of evolutionary pressures

Evolutionary pressures were evaluated both by variant status and glioma subtype using the neutralitytestr algorithm as previously described (R package: neutralitytestr v.0.0.2, <https://github.com/marcjwilliams1/neutralitytestr>)<sup>32</sup>. Individual variant allele frequency vectors were merged at the level of glioma subtype by variant status. Only mutations found in copy-neutral regions were included in these analyses. For all else, default parameters were used. Merged variant allele frequency distributions were deemed to be selected when the neutral null hypothesis was rejected using several metrics. Tests for neutrality required that both  $R^2 < 0.98$  and the area between the two curves of (1) merged variant allele frequency data and (2) a normalized distribution expected under neutrality to be significantly different.

The SubclonalSelection algorithm was applied to GLASS mutation data to measure the selection strength in individual tumour samples (Julia package: SubclonalSelection, <https://github.com/marcjwilliams1/SubClonalSelection.jl>)<sup>16</sup>. Patients that had samples at both time points with a TITAN-defined purity estimate  $\geq 0.5$  and  $\geq 25$  subclonal mutations in diploid regions were included. Mean coverage across all mutations was used as the 'read\_depth' input parameter and the model was run with the recommended  $10^6$  iterations and 1,000 particles. Samples were classified as neutral or selected based on the model that had the highest probability, in line with the prior applications to TCGA data<sup>16</sup>. Classification based on the highest model probability yielded stable results as there was not a significant change in proportions when setting a higher classification probability threshold ( $P > 0.05$ , Pearson's chi-square test, for both probability thresholds of 0.6 and 0.7). At all three probability thresholds (0.5, 0.6 and 0.7), Kaplan–Meier survival analyses between selection at recurrence and overall survival continued to indicate that patients with IDH-wild-type tumours that were selected had a worse overall survival ( $P = 0.03$  ( $n = 81$ ),  $P = 0.01$  ( $n = 66$ ) and  $P = 0.01$  ( $n = 56$ ), respectively).

## Mutation clonality

Each patient's clonal architecture was inferred using PyClone (v.0.13.1) by grouping SNVs into clonal clusters (<https://github.com/aroth85/pyclone>)<sup>60</sup>. The patient-level input mutation matrix was reduced by limiting to sites with at least  $30\times$  coverage across all samples. PyClone was subsequently run using a binomial density model, connected initiation, and 10,000 iterations. Sample purities were provided for each patient and parental copy number (minor and major allele counts) from TITAN were given. PyClone results were post-processed using a burn-in of 1,000, thin of 1, minimum cluster size of 2 and a maximum number of clusters per patient of 12. Individual mutations were determined to be clonal if the PyClone CCF values were  $\geq 0.5$ , subclonal for mutations with  $CCF \geq 0.1$  and  $CCF < 0.5$ , mutations were considered non-clonal when  $CCF < 0.1$ , as previously described<sup>61</sup>.

## CNV clonality

Allele-specific copy number, tumour purity and ploidy estimates were derived using a probabilistic model (TITAN, v.1.19.1) for both whole-genome and whole-exome sequencing samples<sup>62</sup>. TITAN was supplied with the tumour denoised read counts output by GATK DenoiseReadCounts and the tumour allelic counts at loci found to be heterozygous in control samples output by ModelSegments. An 'alphaK' (and 'alphaKHigh') parameter of 2,500 and 10,000 was used for exomes and genomes, respectively. The patient sex was provided to improve fitting allosomes. For each tumour–control pair, TITAN was run assuming an initial ploidy of two or three, and assuming one to three clusters, resulting in a total of six possible solutions per tumour/control pair. To select the optimal solution, TITAN's internal select-Solution function was used with a threshold of 0.15 giving additional weight to diploid solutions.

## Timing analysis

The CCF values output by TITAN or PyClone were used for separately timing copy number changes or mutations. To time specific copy number changes in genes, the average CCF for that gene was calculated. When timing mutations in genes, the highest CCF amongst the non-synonymous mutations was taken.

## Neoantigen analyses

Neoantigens in this analysis were defined as all 8–11-mer peptides that arose from an exonic nonsynonymous SNV or indel and bound their respective patient's HLA class I molecules at a binding affinity score (half-maximal inhibitory concentration,  $IC_{50}$ ) that was  $\leq 500$  nM and better than or equal to the wild-type form of the peptide. Each patient's four-digit HLA class I types were inferred using OptiType (v.1.3.1, <https://github.com/FRED-2/OptiType>) run on each patient's matched normal sample<sup>63</sup>. VCF files for each tumour sample were annotated using Variant Effect Predictor (ensembl) with the 'downstream' and 'wildtype' plugins. Neoantigens from these VCFs were then called using pVACseq (v.4.0.10, <https://github.com/griffithlab/pVAC-Seq>)<sup>41</sup> run using netMHCpan (v.2.8, <http://www.cbs.dtu.dk/services/NetMHCpan-2.8/>)<sup>64</sup>. For each pVACseq run, epitope length was set to 8, 9, 10 or 11, minimum binding affinity fold change was set to 1, and downstream sequence length was set to full, with default parameters used for all other settings.

Downstream neoantigen analyses were performed using the pVACseq output linked to its respective mutation information. Neoantigen-causing mutations were defined as all mutations that gave rise to at least one neoantigen. The observed-to-expected neoantigen ratio was calculated using a previously developed approach that compares each tumour's observed neoantigen rate to an empirically derived expected rate that assumes no selection against neoantigen-causing mutations<sup>39</sup>. From the gold set samples in the GLASS cohort ( $n = 222$ ), define  $\bar{N}_s$  to be the expected number of nonsynonymous missense SNVs per synonymous SNV with trinucleotide context  $s$ .  $\bar{B}_s$  is then defined as the expected number of neoantigen-generating missense SNVs per nonsynonymous missense SNV with trinucleotide context  $s$ . For a given sample  $i$ , define  $Y_i$  as the sample's set of synonymous SNVs and  $s(m)$  to be a synonymous SNV with trinucleotide context  $m$ . The expected number of nonsynonymous missense SNVs,  $N_{pred,i}$ , and neoantigen-causing mutations,  $B_{pred,i}$ , can then be calculated as follows:

$$N_{pred,i} = \sum_{m \in Y_i} \bar{N}_{s(m)}$$

$$B_{pred,i} = \sum_{m \in Y_i} \bar{N}_{s(m)} \bar{B}_{s(m)}$$

To obtain the final neoantigen depletion ratio,  $R_i$ , of sample  $i$ , the observed number of neoantigen-causing mutations in the sample,  $B_{obs,i}$ , is divided by the sample's observed number of nonsynonymous missense SNVs,  $N_{obs,i}$ , and then this ratio is divided by the ratio of  $B_{pred,i}$  and  $N_{pred,i}$ . Thus:

$$R_i = \frac{B_{obs,i}/N_{obs,i}}{B_{pred,i}/N_{pred,i}}$$

For analyses examining clonal/subclonal neoantigen ratios, the observed and expected numbers were calculated by subsetting the SNVs of a sample by the respective criteria and then recalculating the ratio as described above. To mitigate overfitting, all analyses presented here used samples from patients with at least three neoantigen-causing mutations in their primary and recurrent tumours.

## Immune cell analyses

CIBERSORT relative immune cell fraction data used in downstream neoantigen analyses were downloaded from a previous publication<sup>38</sup>.

## Statistical methods

All data analyses were conducted in R 3.4.2, Python 2.7.15, PostgreSQL 10.5, and Julia 0.7. All survival analyses including Kaplan–Meier plots and Cox proportional hazards models were conducted using the R packages survival and survminer.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

All de-identified, non-protected access somatic variant profiles and clinical data are accessible via Synapse (<http://synapse.org/glass>). Raw data of the various sequencing datasets can be obtained in the Supplementary Information.

## Code availability

All custom scripts and pipelines are available on the project's github page (<https://github.com/TheJacksonLaboratory/GLASS>).

48. Brennan, C. W. et al. The somatic genomic landscape of glioblastoma. *Cell* **155**, 462–477 (2013).
49. Droop, A. et al. How to analyse the spatiotemporal tumour samples needed to investigate cancer evolution: a case study using paired primary and recurrent glioblastoma. *Int. J. Cancer* **142**, 1620–1626 (2018).
50. Mazar, T. et al. DNA methylation and somatic mutations converge on the cell cycle and define similar evolutionary histories in brain tumors. *Cancer Cell* **28**, 307–317 (2015).
51. Kim, J. et al. Spatiotemporal evolution of the primary glioblastoma genome. *Cancer Cell* **28**, 318–328 (2015).
52. Suzuki, H. et al. Mutational landscape and clonal architecture in grade II and III gliomas. *Nat. Genet.* **47**, 458–468 (2015).
53. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **34**, 3600 (2018).
54. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics*. **43**, 11.10.11–11.10.33 (2013).
55. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
56. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
57. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
58. Beroukhi, R. et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl Acad. Sci. USA* **104**, 20007–20012 (2007).
59. Taylor, A. M. et al. Genomic and functional approaches to understanding cancer aneuploidy. *Cancer Cell* **33**, 676–689.e673 (2018).
60. Roth, A. et al. PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11**, 396–398 (2014).
61. Turajlic, S. et al. Tracking cancer evolution reveals constrained routes to metastases: TRACERx Renal. *Cell* **173**, 581–594.e512 (2018).
62. Ha, G. et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.* **24**, 1881–1893 (2014).
63. Szolek, A. et al. OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* **30**, 3310–3316 (2014).
64. Hoof, I. et al. NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* **61**, 1–13 (2009).

**Acknowledgements** This work is made possible by the patients and their families whom generously contributed to this study. This work is supported by the National Brain Tumor Society, Oligo Research Fund; Cancer Center Support grants P30CA16672 and P30CA034196;

Cancer Prevention & Research Institute of Texas (CPRIT) grant number R140606; Agilent Technologies (R.G.W.V.); the National Institutes of Health–National Cancer Institute for the following grants: NCI CA170278 (L.M.P., T.M.M., H.N.), NCI R01CA222146 (L.M.P., H.K.N.), NCI R01CA230031 (J.H.C., J.N.), NCI R01CA188288 (J.S.B.-S., R.B., P.B., K.L.L., A. Chakravarty, A.E.S.), R01CA179044 (A. Iavarone), U54CA193313 (A. Iavarone). The National Brain Tumor Society (W.K.A.Y., J.F.d.G.). Brain Tumour Northwest tissue bank (including the Walton research tissue bank) is supported by the Sidney Driscoll Neuroscience Foundation and part of the Walton Centre and Lancashire Teaching Hospitals NHS Foundation Trusts (A.F.B., M.D.J.). This work was supported by a generous gift from the Dabbiere family (J.F.C.). Support is also provided by a Leeds Charitable Foundation grant (9R11/14-11 to L.F.S.), University of Leeds Academic Fellowship (11001061) (L.F.S.) and Studentship (11061191) (G. Tanner) as well as Leeds Teaching Hospitals NHS Trust (A. Chakravarti, A. Ismail). The Leeds Multidisciplinary Research Tissue Bank staff was funded by the PPR Foundation and The University of Leeds (S.C.S.). Funds were received from The Brain Tumour Charity (C.W., grants 10/136 & GN-000580, B.A.W., 200450). Ghazaleh Tabatabai is funded by EKFS 2015\_Kolleg\_14. R01CA218144 (P.S.L., E.J.C., J.C., A.K.L.) and Strain for the Brain, Milwaukee, WI (P.S.L., E.J.C., J.C., A.K.L.). E.K. is recipient of an MD-Fellowship by the Boehringer Ingelheim Fonds and is supported by the German National Academic Foundation. The Leeds Multidisciplinary Research Tissue Bank staff was funded by the PPR Foundation and part of the University of Leeds (S.C.S.). GLASS-Austria was funded by the Austrian Science Fund project KLI394 (A.W.). GLASS-Germany was funded by the German Ministry of Education and Research (BMBF) 031A425 (G. Reifemberger, P.L.) and German Cancer Aid (DKH) 70-3163-Wi 3 (M.W.). GLASS-NL receives support from KWF/Dutch Cancer Society project 11026 (M.C.M.K., P.W., R.G.W.V., P.J.F., J.M.N., M. Smits, B.A.W.). We thank the University of Colorado Denver Central Nervous System Biorepository (D.R.O.) for providing tissue samples. Sponsoring was also received from the National Institute of Neurological Disorders and Stroke (NINDS R01NS094615, G. Rao), F.S.V. is supported by a postdoctoral fellowship from The Jane Coffin Childs Memorial Fund for Medical Research. F.P.B. is supported by the JAX Scholar program and the National Cancer Institute (K99 CA226387); K.C.J. is the recipient of an American Cancer Society Fellowship (130984-PF-17-141-01-DMC). We thank the Jackson Laboratory Clinical and Translation Support team for coordinating all data transfer agreements. We thank M. Wimsatt for assistance in graphic design.

**Author contributions** Sequencing data coordination was performed by H.K., F.P.B. and K.C.J., and clinical data coordination was by A.D.M. and O.A. Data analysis was led by F.P.B. and K.C.J. in collaboration with K.J.A., S.B.A., J.H.C., H.K., E.K., J.N., L.F.S., G. Tanner, F.S.V. and R.G.W.V. Clinical analysis was performed by F.P.B., K.C.J., A.D.M., L.M.P. and C.W. Pathology review was completed, in part, by A. Chakravarty, J.T.H., A. Ismail, A.W., H.K.N., K.L.L., G. Reifemberger and K.A. F.P.B., K.C.J., A.D.M., F.S.V. and R.G.W.V. wrote the manuscript. K.D.A., J.H. and J.F.d.G. coordinated the GLASS-MDACC cohort. L.F.S. was the lead coordinator of the GLASS-Leeds cohort and B.A.W. the lead coordinator of GLASS-Netherlands. D.M.A., D.A., P.B., J.S.B.-S., R.B., C.B., P.K.B., D.J.B., A.R.B. A. Chakravarty, A. Chakravarti, E.J.C., J.F.C., G.F., M.N.F., A. Iavarone, M.D.J., M.K., P.S.L., M.L., P.L., K.L.L., T.M.M., T.M., A.M.M., D.-H.N., N.N., H.K.N., C.Y.N., S.P.N., H.N., D.R.O., C.-K.P., L.M.P., G. Rao, B.R., J.K.S., S.C.S., A.E.S., M. Schuster, L.F.S., H.S., E.G.V.M., C.W., M.W., G.W. and A.W. contributed to sample acquisition and processing. All co-authors including K.A., P.B., A.F.B., K.R.B., E.B.C., J.C., P.J.F., H.K.G., M. R. Grimmer, P.V.G., M. R. Gilbert, A.K.L., K.L.M., J.M.N., R.R., G. Reifemberger, B.L.S., P.A.S.S., M. Smits, G. Tabatabai, P.W., W.K.A.Y. and G.Z. discussed the results and commented on the manuscript and Supplementary Information. R.G.W.V. was the project lead and coordinator.

**Competing interests** R.G.W.V. declares equity in Boundless Bio, Inc. M.K. receives research grants from BMS and ABBVie. P.K.B. is a consultant for Lilly, Genentech-Roche, Angiochem and Tesaro. P.K.B. receives institutional funding from Merck and Pfizer and honoraria from Merck and Genentech-Roche. W.K.A.Y. serves in a consulting or advisory role at DNATRIX Therapeutics. M.W. receives funding from Acceleron, Actelion, Bayer, Isarna, Merck, Sharp & Dohme, Merck (EMD, Darmstadt), Novocure, OGD2, Pigur and Roche as well as honoraria from BMS, Celldex, Immunocellular Therapeutics, Isarna, Magforce, Merck, Sharp & Dohme, Merck (EMD, Darmstadt), Northwest Biotherapeutics, Novocure, Pfizer, Roche, Teva and Tocagen. G. Reifemberger receives funding from Roche and Merck (EMD, Darmstadt) as well as honoraria from AbbVie. M. Smits is a central reviewer for Parexel Ltd and honoraria are paid to the institution. G. Tabatabai reports personal fees from Bristol-Myers-Squibb, personal fees from AbbVie, personal fees from Novocure, personal fees from Medac, travel grants from Bristol-Myers-Squibb, education grants from Novocure, research grants from Roche Diagnostics, research grants from Medac, membership in the National Steering board of the TIGER NIS (Novocure) and the International Steering board of the ON-TRK NIS (Bayer).

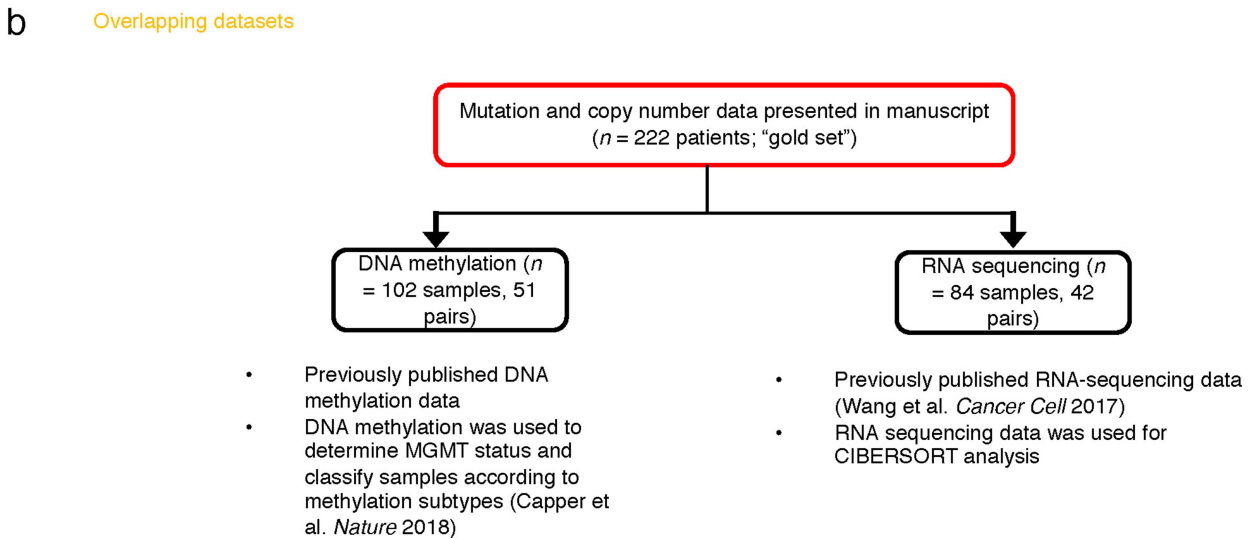
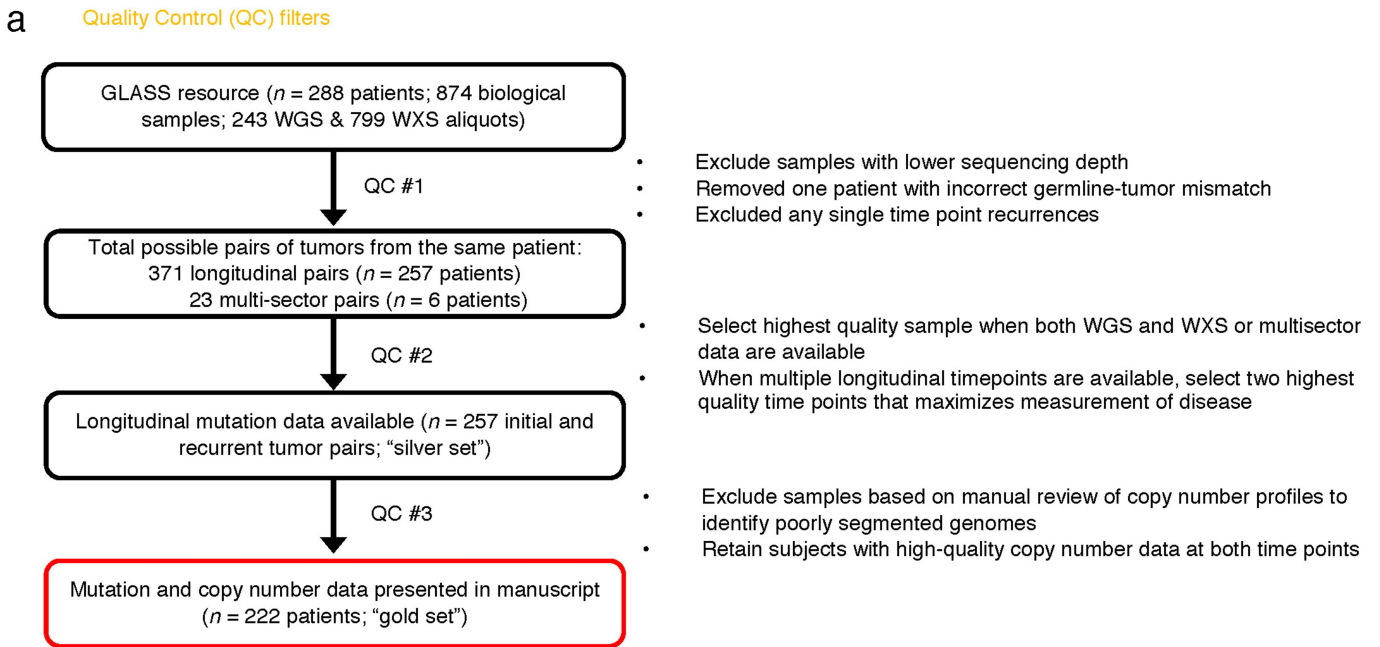
## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1775-1>.

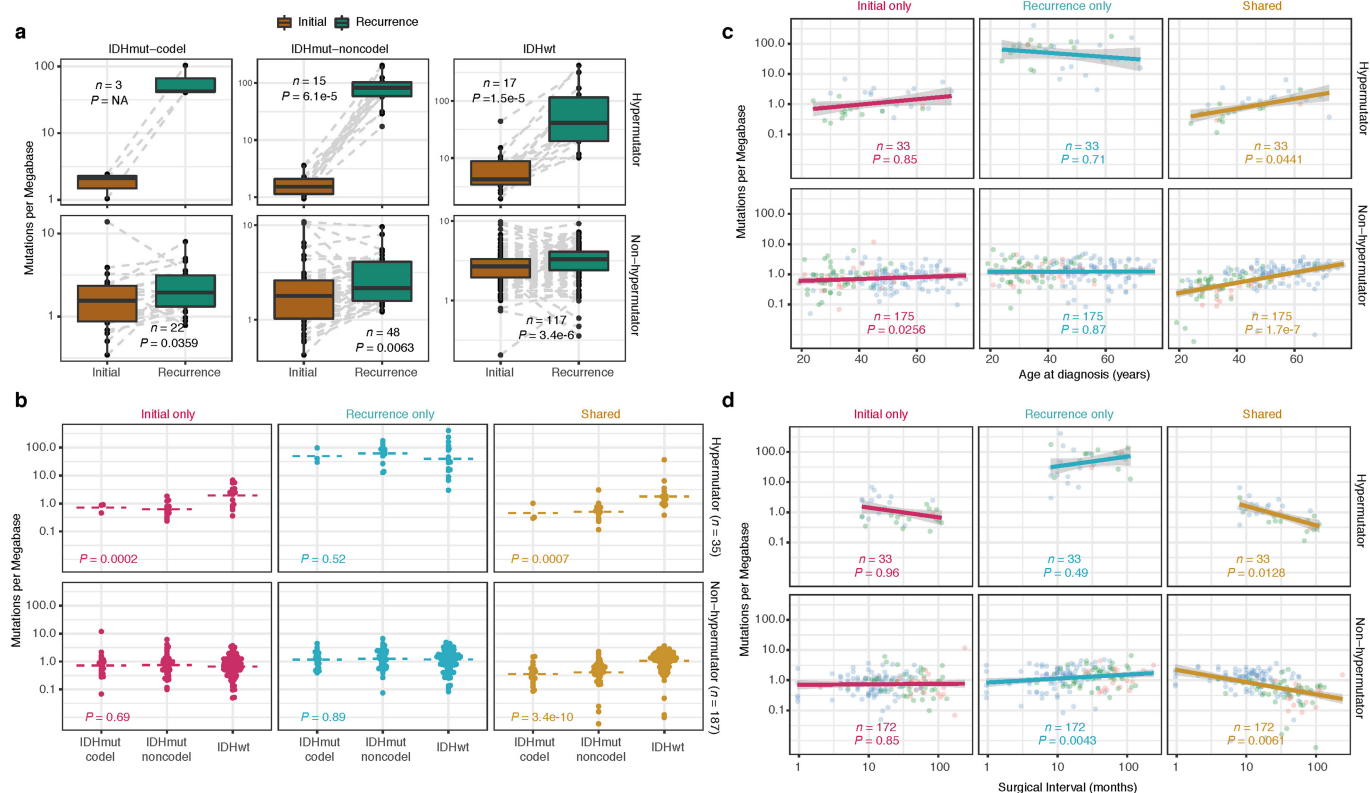
**Correspondence and requests for materials** should be addressed to R.G.W.V.

**Peer review information** Nature thanks Kamila Naxerova, Wolfgang Wick and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

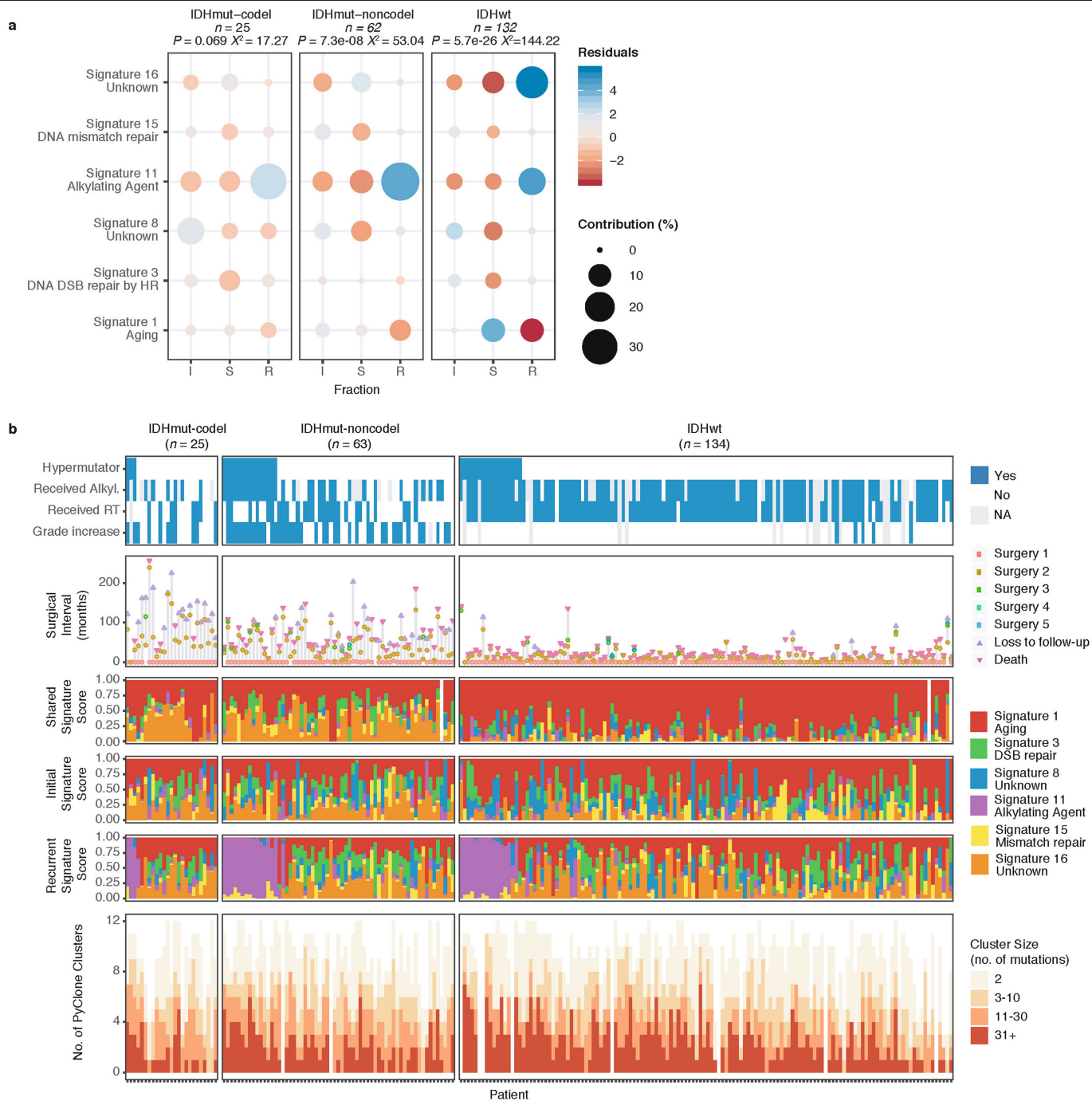


**Extended Data Fig. 1 | Sample selection.** **a**, Quality control workflow steps identifying all GLASS samples available as a resource and the identification of the highest quality set of patient pairs ( $n = 222$ ) used for the presented mutational and copy number analyses. **b**, Additional available datasets.



**Extended Data Fig. 2 | Mutation burden by time point and subtype. a,** Box plots and paired lines depicting coverage-adjusted mutation frequencies in initial and matched recurrent samples across three subtypes. Wilcoxon signed-rank test  $P$  values and sample sizes are indicated. **b,** Bee swarm plot depicting coverage-adjusted mutation frequencies in fractions by subtype. Dashed line indicates the mean.  $P$  values comparing three subtypes were determined by

one-way analysis of variance (ANOVA). **c,** Scatter plot showing the relationship between age at diagnosis and coverage adjusted mutation burdens by subtype and fraction.  $P$  values were determined by the linear model and adjusted by subtype. **d,** Similar to the analysis in **c** but showing the relationship between time to recurrence and coverage-adjusted mutation burdens.

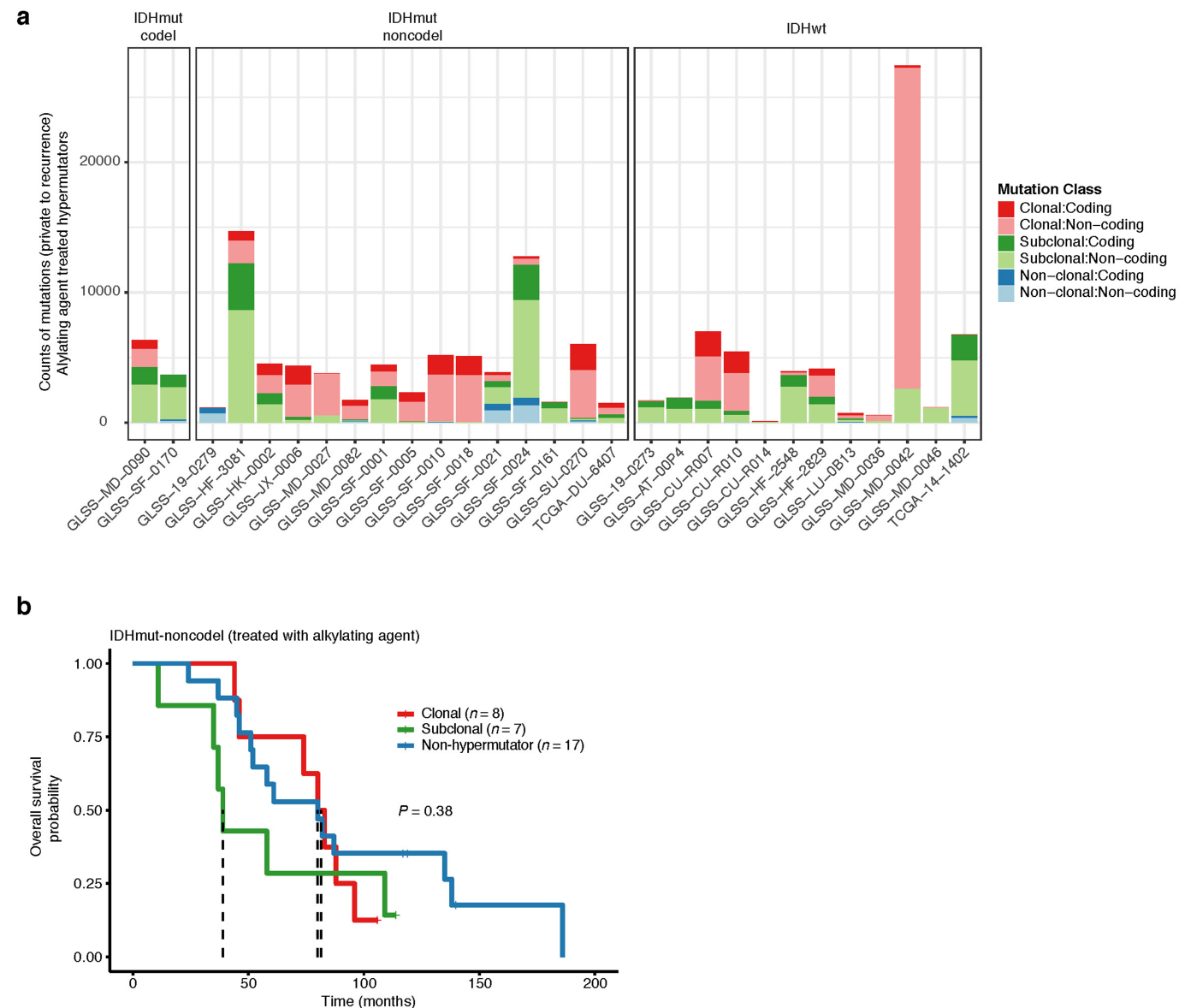


**Extended Data Fig. 3 | Mutational signatures by fraction and subtype.**

**a.** Correlation plot showing the Pearson's chi-squared ( $\chi^2$ ) residuals for each signature by fraction and subtype. A  $\chi^2$  test was performed for each subtype and  $P$  values are indicated. Positive residuals (blue) indicate a positive correlation, whereas negative residuals (red) indicate an anti-correlation. The

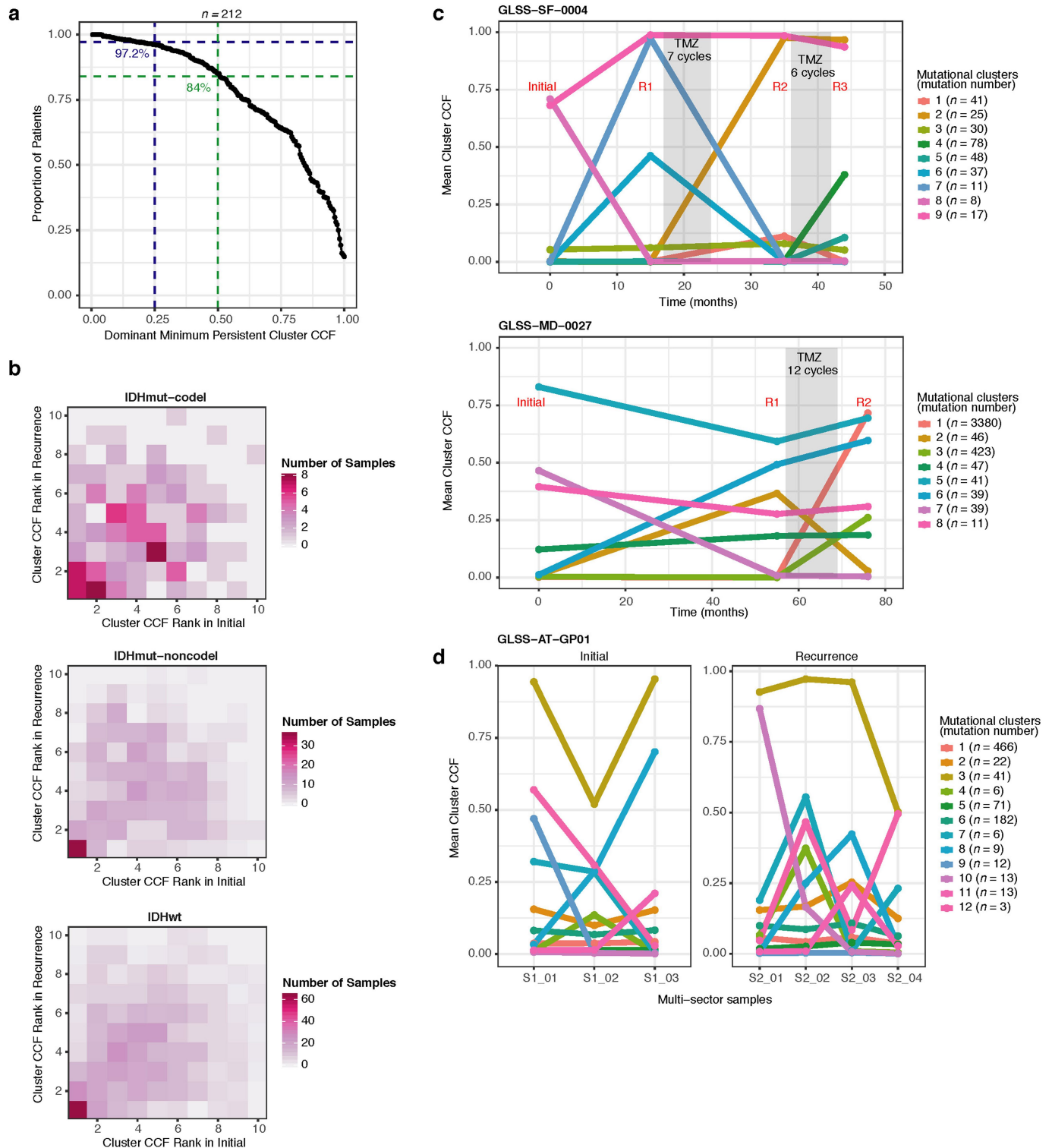
point size reflects the contribution to the  $\chi^2$  estimate. **b.** Patients were ordered as in Fig. 1a, and relevant clinical information is provided alongside the fraction-specific mutational signatures. PyClone mutational clusters are also presented.





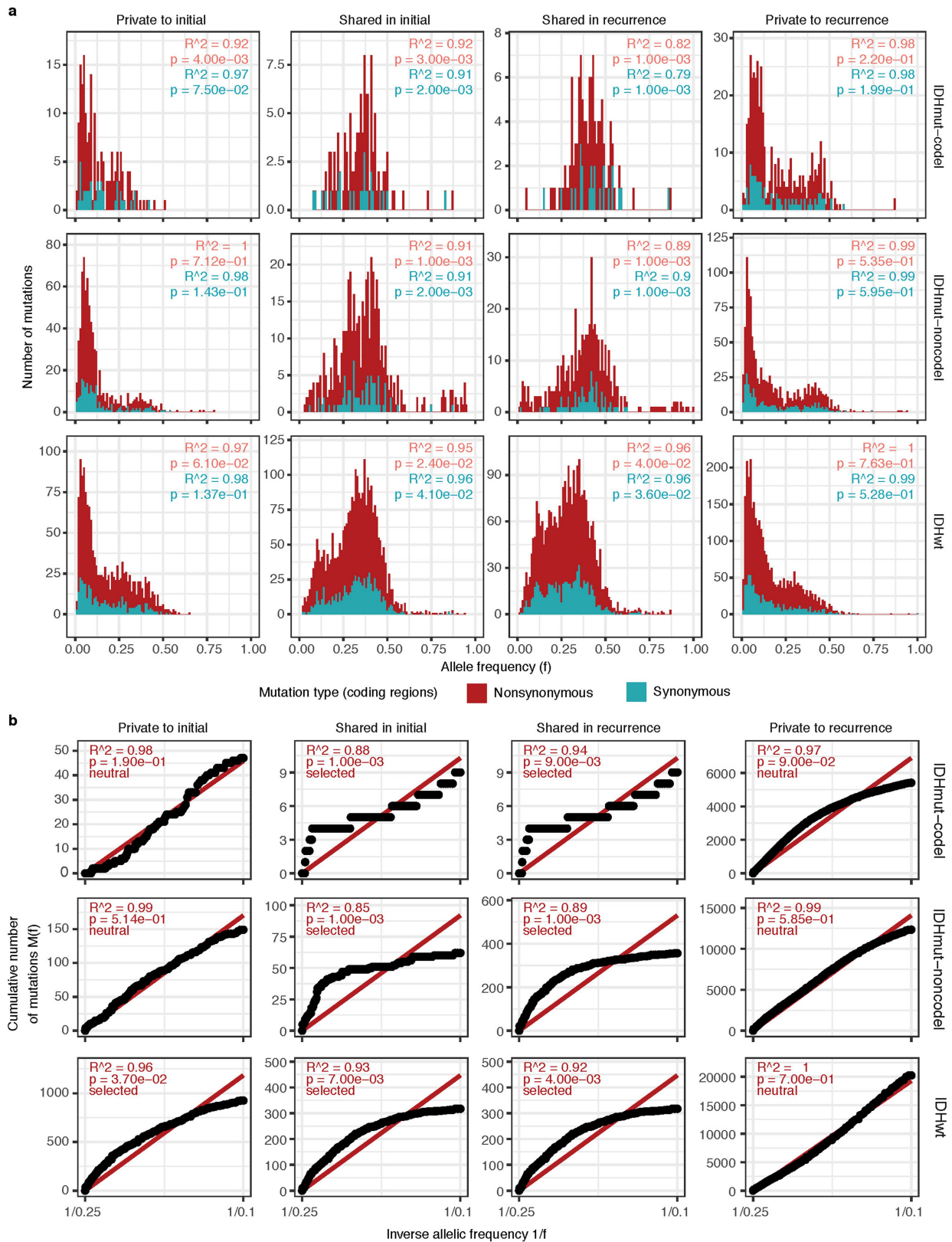
**Extended Data Fig. 4 | Hypermutator clonality. a**, Bar plots represent counts of recurrence-only mutations per hypermutator tumour that were known to receive treatment alkylating agent and were successfully run through the PyClone algorithm. Colours indicate mutation clonality and colour intensity indicates whether the mutations resulted in coding changes. **b**, Kaplan-Meier

curve comparing the survival of alkylating agent-treated IDH-mutant-noncodon hypermutator tumours that were predominantly clonal ( $n=8$ ), predominantly subclonal ( $n=7$ ) or non-hypermutator ( $n=17$ ). Limited to tumours with available PyClone data.  $P$  value determined by log-rank test.



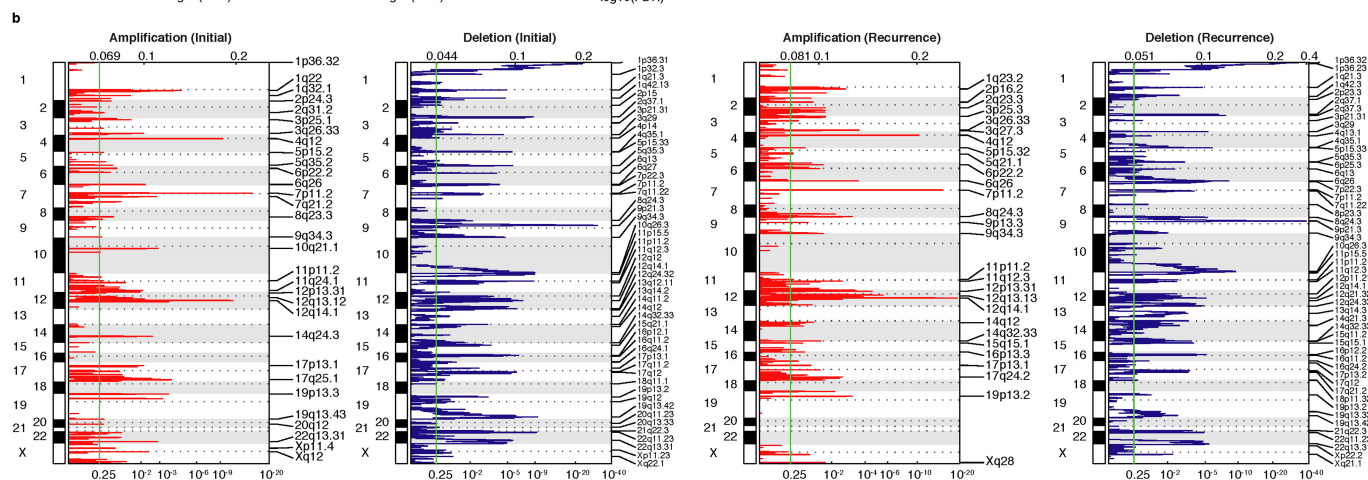
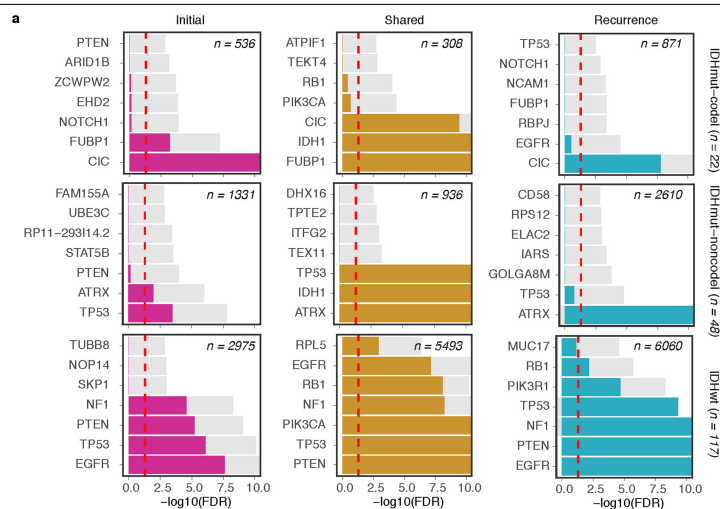
**Extended Data Fig. 5 | Clonal structure evolution over time.** **a**, The minimum CCF of the most persistent (shared between initial and recurrence) PyClone cluster. **b**, Comparison of PyClone clusters ranked by CCF in matched initial and recurrent tumours, as in Fig. 2b, but separated by subtype. **c**, **d**, Examples of

cluster CCF dynamics over time in three separate samples, including two multi-time point samples (**c**) and one multi-sector sample (**d**). These additional data are available in the GLASS resource, but only two time-separated samples were used throughout to ensure clarity.



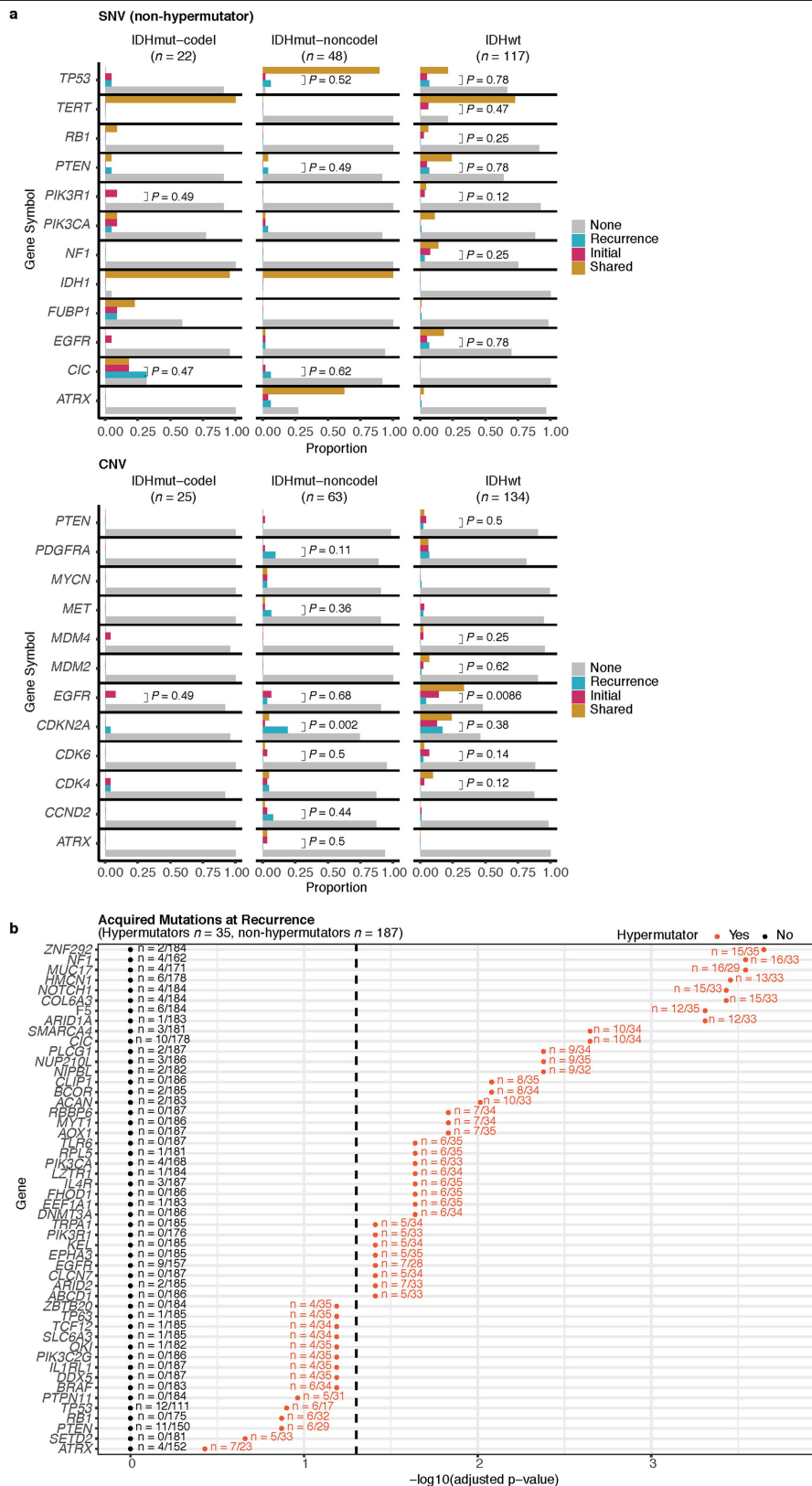
**Extended Data Fig. 6 | Distribution of variant allele fraction. a**, Distributions of non-hypermutator variant allele fraction for copy-neutral variants in coding regions ( $n=181$  patients). Variants are separated by subtype, fraction and also the variant was non-synonymous or synonymous mutation in a coding region.  $R^2$  goodness-of-fit measure and associated  $P$  values are shown. Note that these

data consider only the coding portion of genome, whereas Fig. 2d presents both coding and non-coding data. **b**, The cumulative distribution of the subclonal mutations in copy-neutral regions for hypermutators ( $n=31$  patients). For each variant fraction and subtype, the  $R^2$  goodness-of-fit measure and  $P$  values are shown.



**Extended Data Fig. 7 | Driver gene nomination. a**, Local (gene-wise) dN/dS estimates by subtype (rows) and fraction (columns). Genes are sorted by  $Q$  value and  $P$  value. The  $Q$  value is shown in colour, whereas the  $P$  value is indicated in light grey. The  $Q$  value threshold of 0.05 is indicated by a horizontal

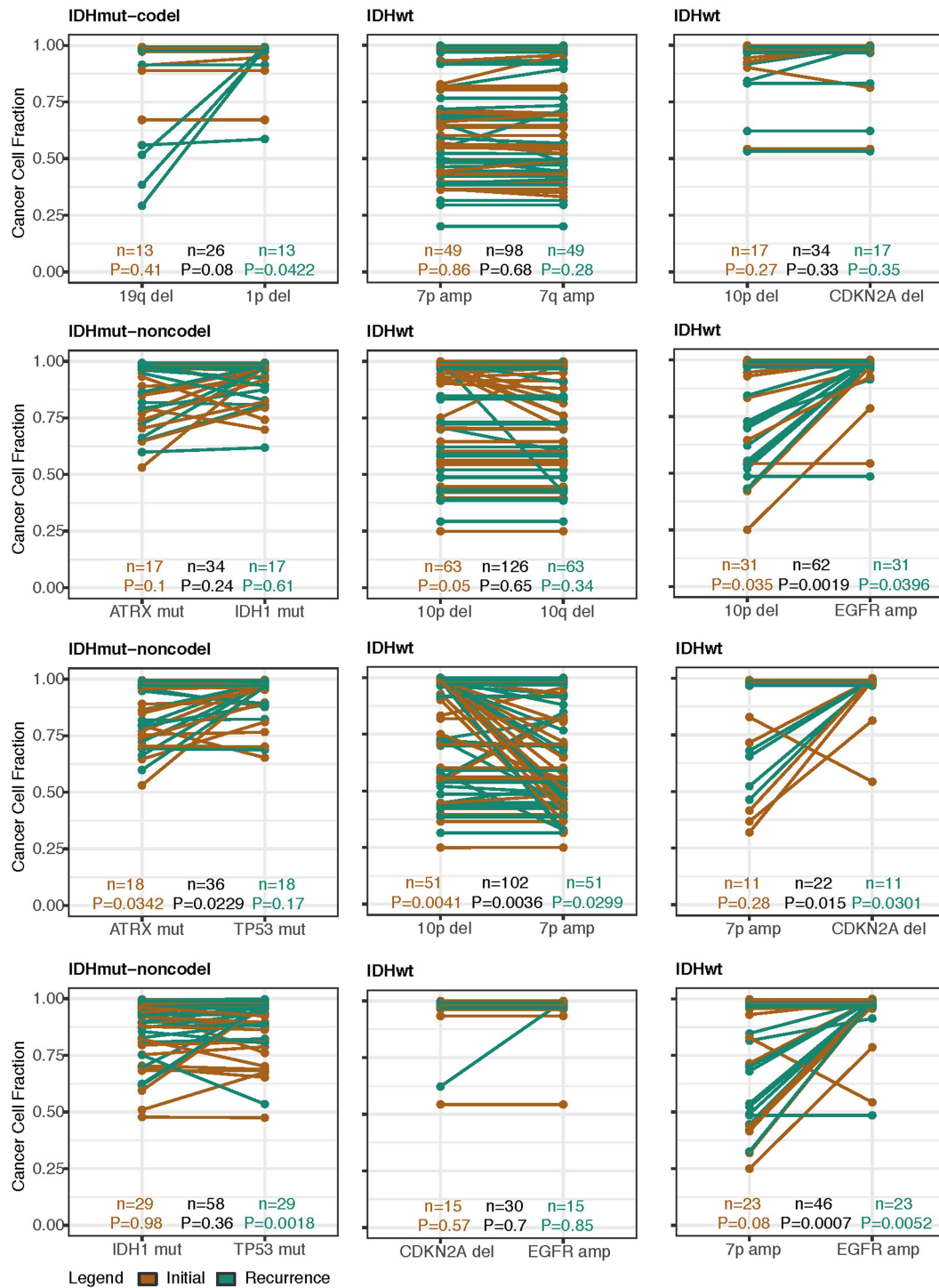
red line. **b**, GISTIC significant amplification (red) and deletion (blue) plots in initial (left) and recurrent tumours (right). Chromosomal locations are ordered on the y-axis,  $Q$  values are shown on the x-axis, and selected drivers are indicated by their chromosomal location on the right.



**Extended Data Fig. 8 | Driver acquisition over time. a**, Tabulated numbers of SNV (top) and CNV (bottom) driver events that were shared, initial-only or recurrence-only.  $P$  values were determined by a two-sided Fisher test comparing the initial-only fraction to the recurrence-only fraction testing for acquisition. **b**, One-sided Fisher test comparing the initial-only fraction to the

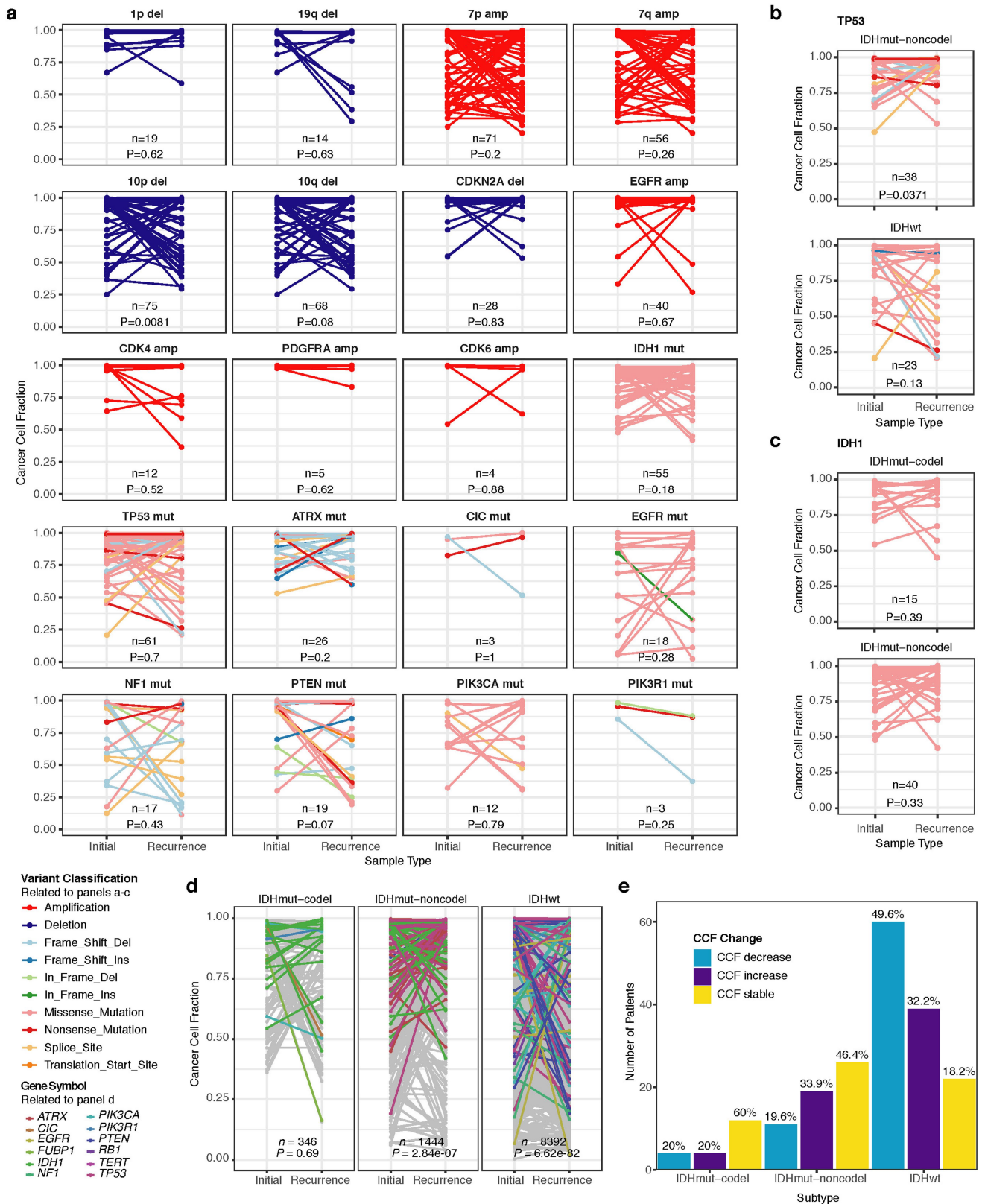
recurrence-only fraction among previously implicated glioma drivers testing for driver acquisition.  $P$  values were adjusted for multiple testing using the false discovery rate (x axis). Hypermutators (red) and non-hypermutators (black) were separately analysed.





**Extended Data Fig. 9 | Intra-tumour CCF comparison.** Ladder plots comparing the CCF of co-occurring drivers in single tumour samples. The colour of the lines and points indicates whether the sample shown is an initial

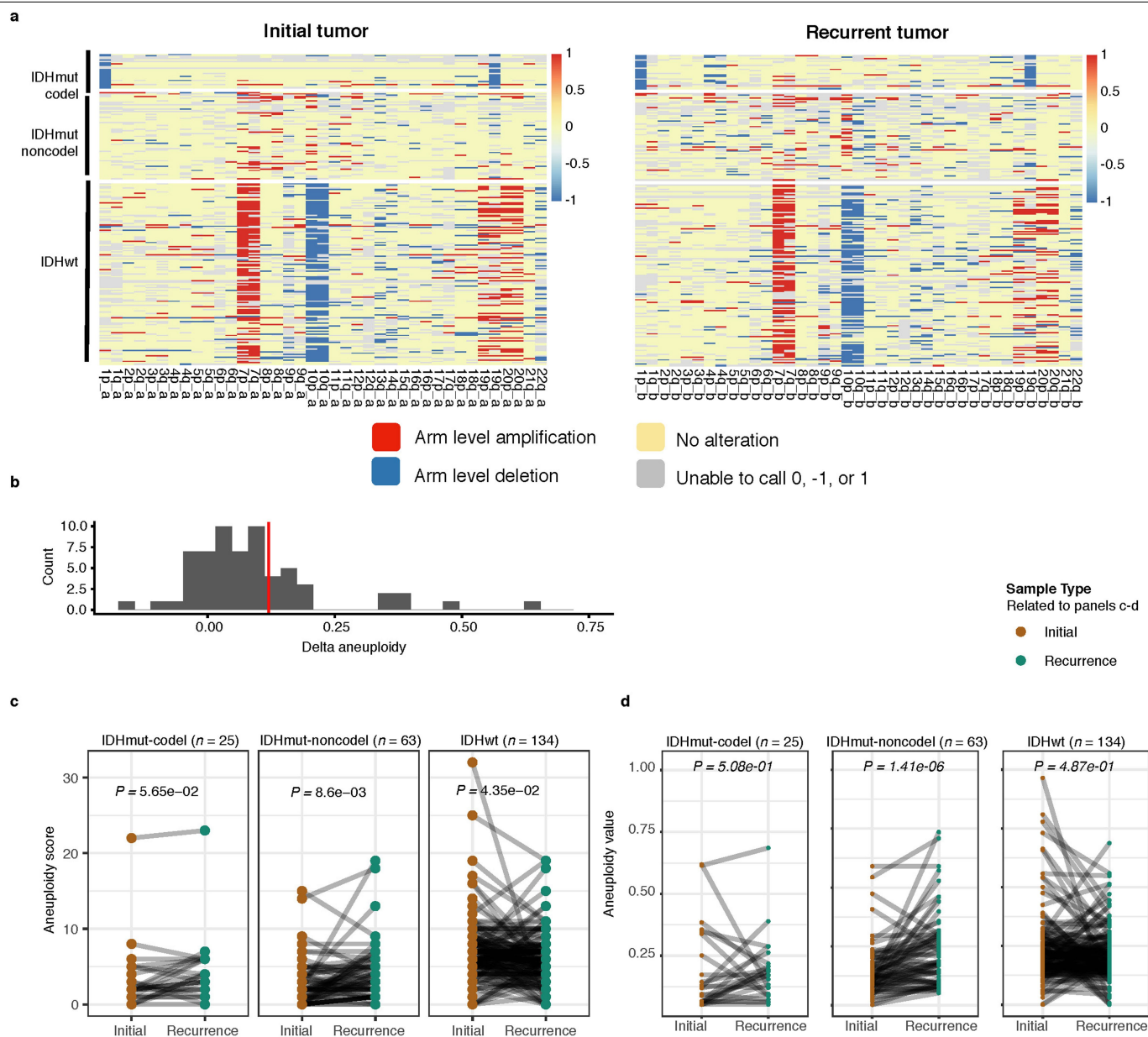
(brown) or recurrent (green) tumour. *P* values determined by two-sided Wilcoxon rank-sum test for all initial samples, recurrent samples, as well as all samples (black).



**Extended Data Fig. 10 | Between time point intra-patient CCF comparison.**

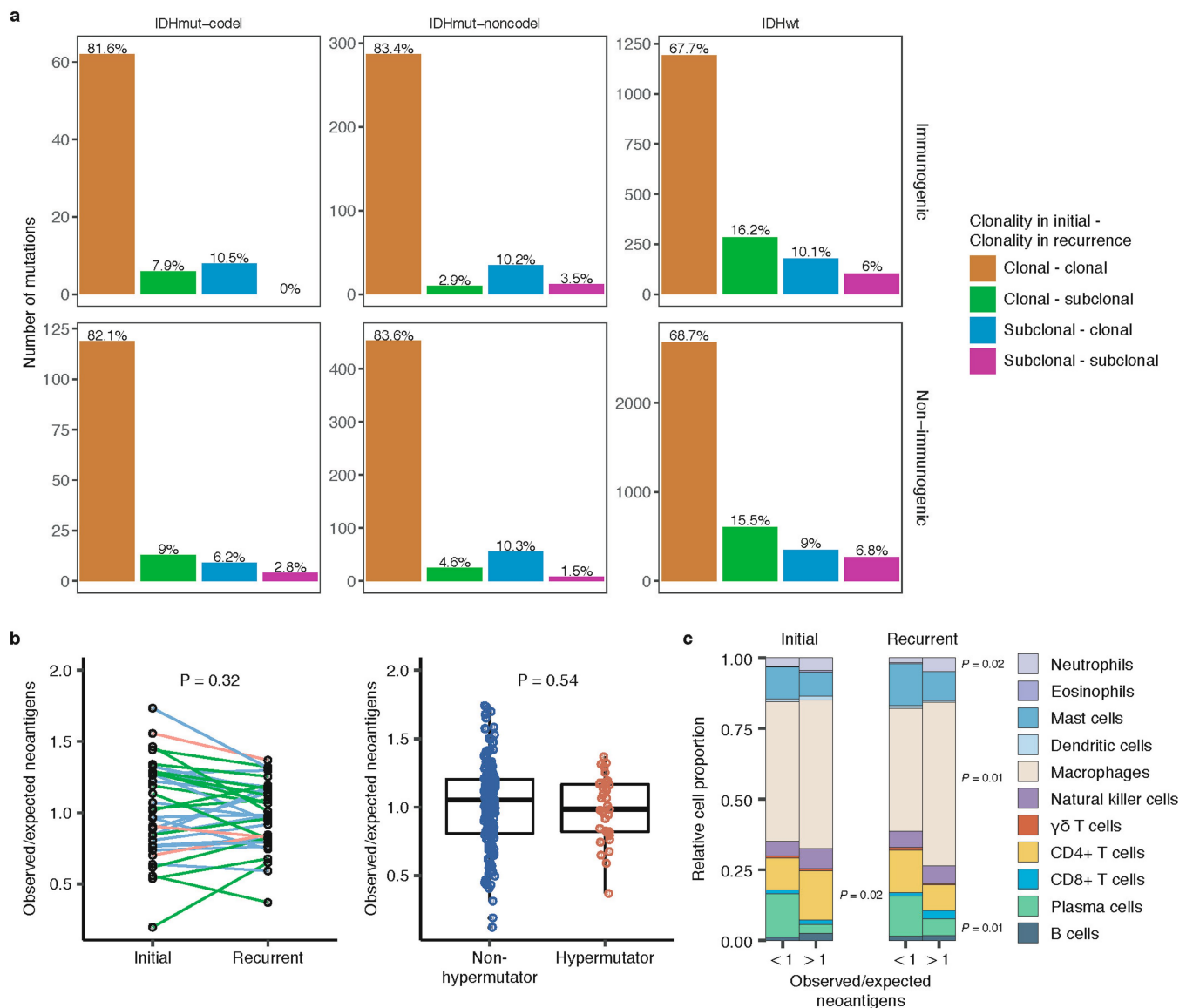
**a**, Driver gene CCF comparison between initial and matched recurrences. Lines are coloured by variant classification. *P* values determined by two-sided Wilcoxon rank-sum test. **b**, TP53 CCF by subtype, otherwise as in **a**. **c**, IDH1 CCF by subtype, otherwise as in **a**. **d**, Ladder plot visualizing change in CCF across all SNVs between initial and recurrent tumours, separated by subtype. *P* values

determined by Wilcoxon rank-sum test. **e**, Initial and recurrent mutations in each patient were compared using a Wilcoxon rank-sum test. Bar plot with counts of patients in each subtype are shown. Patients lacking significant change are shown in yellow, and those with a significant increase or decrease are shown in dark and light blue, respectively.



**Extended Data Fig. 11 | Aneuploidy calculation.** **a**, Heat map displaying the chromosomal arm-level events (x axis) with patients represented in each row. Patients are placed in the same order for both the initial (left) and recurrence (right). White space was inserted as a break between the three subtypes.

**b**, Distribution of total aneuploidy difference. Acquired aneuploidy determination (upper-quartile) indicated with a red line. **c**, Comparison of aneuploidy score between initial and recurrent tumours separated by subtype **d**. As in **c**, comparing aneuploidy value.



**Extended Data Fig. 12 | Neoantigen evolution and cellular analysis. a,** Bar plots representing the number of shared mutations that give rise to neoantigens (top row, 'immunogenic') and those that do not give rise to neoantigens (bottom row, 'non-immunogenic') stratified by longitudinal clonality ('(clonality in initial) - (clonality in recurrence)') and further separated by subtype. The percentage of longitudinal clonality per subtype and mutation is shown. **b,** Left, ladder plot depicting the difference in observed-to-expected neoantigen ratio between the initial and recurrent tumours of patients with hypermutated tumours at recurrence. Each set of points connected by a line represents one tumour ( $n = 70$ ). Right, box plot depicting the distribution of observed-to-expected neoantigen ratios in

recurrent tumours stratified by hypermutator status ( $n = 35$  and  $183$  for hypermutators and non-hypermutators, respectively). Each box spans quartiles, with the lines representing the median ratio for each group. Whiskers represent absolute range, excluding outliers.  $P$  values were determined by a paired and an unpaired two-sided  $t$ -test, for left and right graphs, respectively. **c,** Stacked bar plots depicting the average relative fraction of 11 CIBERSORT cell types in the neoantigen depleted (<1) and non-depleted (>1) initial and recurrent tumour subgroups.  $P$  values to the right of each plot indicate a significant difference between the depleted and non-depleted groups for the noted cell type at that time.

# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- |                                     |  |
|-------------------------------------|--|
| n/a                                 | Confirmed  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The exact sample size ( <i>n</i> ) for each experimental group/condition, given as a discrete number and unit of measurement   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of all covariates tested   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i> ) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted<br><i>Give P values as exact values whenever suitable.</i>                     |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i> ), indicating how they were calculated   |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

## Software and code

Policy information about [availability of computer code](#)

Data collection	No software was used for data collection.
Data analysis	<p>MultiQC version: 1.6a0 (quality assessment)  <b>FastQC 0.11.7 (quality assessment)</b>            BWA MEM 0.7.17 (alignment)            R 3.4.2 (general data analyses)            Python 2.7.15 (general data analysis)            Julia 0.7 (general data analysis)            PostgreSQL 10.5 (data management)            BCFTools 1.9 (normalize, sort and index variants)            snakemake 5.2.2 (pipeline development)            GATK (including Mutect2) version: 4.1.0.0 (SNV/CNV detection)            freebayes version: 1.2.0 (variant filtering)            vcf2maf version: 1.6.16 (variant filtering and annotation)            MutationalPatterns version: 1.6.1 (mutational signatures)            TITAN version: 1.19.1 (purity, ploidy, CNV clonality estimates)            dndscv (R package) version: 0.0.1.0 (selection strength, nominate driver genes)            alluvial (R package) version: 0.1-2 (visualize longitudinal neutrality)            DBI (R package) version: 1.0.0 (database management)            tidyverse (R package) version: 1.2.1 (data analysis and visualization)            survival (R package) version: 2.42-6 (survival analyses)            neutralitytestr version: 0.0.2 (subtype-level, variant-level selection)            SubClonalSelection version: 0.0.0 (sample-level selection)            PyClone version: 0.13.1 (mutational clusters)            Optityper version: 1.3.1 (HLA class types)  <b>Optityper version: 4.0.10 (neoantigen prediction)</b>            netMHCpan version: 2.8 (neoantigen prediction)            All other custom scripts and pipelines are available on the project's github page (<a href="https://github.com/TheJacksonLaboratory/GLASS">https://github.com/TheJacksonLaboratory/GLASS</a>)</p>

For manuscripts utilizing custom algorithms or software that are not widely used by the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.



## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All deidentified, non-protected access somatic variant profiles and clinical data are accessible via Synapse (<http://synapse.org/glass>). A subset of whole genome and whole exome sequencing data has been deposited in the National Center for Biotechnology Information's Sequencing Read Archive and/or the European Genome/Phenome Archive (EGA). Please see Supplementary Table 1 for availability and accession codes.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. <b>Sample size was a function of availability.</b>
Data exclusions	We defined a quality control process to integrate whole exome and whole genome sequencing data collected from multiple cohorts. As shown in Extended Data Fig. 1, two datasets, Silver and Gold, were constructed to be used for each major analysis type, SNV and CNV, respectively. The two criteria used are intended to provide quality classifications for samples across fingerprinting, coverage, copy number variation (CNV) data and clinical annotation. Fingerprinting was performed using CrosscheckFingerprints (Picard), the purpose of this is to check that all of the input files (readgroups, libraries, samples, files) belong to the same patient, to remove duplicated cases, unmatched samples, and samples of poor quality. Any evidence of mismatch rendered the samples "blocked", otherwise the sample was annotated as "allow". To ensure suitable coverage for mutation calling, samples with near 0 mutation frequency as well as those 2 standard deviations below the mean for either WGS or WXS were annotated as "block". Samples were categorized as "allow", "review", or "block". Copy number data were excluded via manual review of all selected copy number solutions. Manual review consisted of identifying whether data had an atypical or noisy segmentation profile. While we recognize that this strategy is not objective it proved to be an effective strategy for identifying poor performing samples. Insufficient signal, noisy signal, TITAN run fail and unexpected genome stability (little to no copy number changes observed suggesting low purity) were the main reasons for sample exclusion or review. Clinical data was another source of sample filtering. Exclusion of samples was mostly related to sample pairs where surgical interval was very short (1-2 months) and thus did not appear to be a true recurrence. Caution should be used when considering whether a sample represents a true recurrence as no standard set time limits exist. Categories for clinical data include "allow", "interval 1 or less months", "interval 2 or less months", "different location" and "surgical indication" (including "further debulking"). Those interested in using the dataset for further analysis are encouraged to make their own judgments on the criteria they select. The Silver set is filtered to include those pairs with no fingerprinting mismatches and sufficient coverage and is made up of 257 pairs. The Gold set contains 222 pairs, which in addition to the previously mentioned criteria also contain acceptable CNV calls in both samples.
Replication	Replication was limited to select patient samples where both whole genome sequencing and whole exome sequencing was available. <b>All attempts at replication were successful.</b>
Randomization	There was no randomization in this study.
Blinding	All patient samples were deidentified and were assigned a study-specific barcode. Blinding was not relevant to our study since there was no randomization of groups.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials &amp; experimental systems

## Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input type="checkbox"/>	<input checked="" type="checkbox"/> Clinical data

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Human research participants

Policy information about [studies involving human research participants](#)

## Population characteristics

The dataset includes 271 sets of at least two time-separated tumor samples and 17 standalone recurrences. The majority of sets contain two tumor samples (n=246, 85%), with 19 (6.6%) three-tumor sample sets, three (1%) four-tumor sample sets, one (0.3%) with a total of five tumor samples and 17 (5.9%) standalone post-treatment tumor samples. Basic clinical information including age (years), gender, overall survival (months), tumor grade, and tumor histology was available for 90% (260/288) of patients and for 92% (536/584) of tumor samples of the dataset.

Temozolomide and radiation treatment information was available for 68% of the cohort (399/584), data on other treatment modalities was available for 119 patients. Median age at diagnosis of GLASS patients in the IDHmut-noncode and IDHmut-code subtypes were both 34 years old and in the IDHwt group age at diagnosis was 53 years old. This is compared with 46 years for IDHmut-codes, 38 years for the IDHmut-noncodes and 59 years in the TCGA cohort respectively. Patients in our dataset were biased toward longer survival as 261 patients were deemed fit for surgical resection or biopsy at recurrence. Median survival for primary glioblastoma patients was 21 months (95% CI 19–23) in the GLASS cohort versus 15 months in historical cohorts. Patients in this cohort were predominantly treated at teaching/academic centers, which have been shown to be an independent predictive factor of longer survival compared with non-teaching/community hospital settings

All other relevant patient demographics for the GLASS cohort are presented in the Supplement.

## Recruitment

Informed consent was obtained from all study subjects as part of each institution's individual IRB.

## Ethics oversight

All tissue source centers listed in **Supplementary Table 1** obtained study approval by the corresponding institutional review board (IRB) and informed consent from all patients in the cohort. Data pooling at the Jackson Laboratory was performed under the oversight of the IRB at the Jackson Laboratory.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

## Clinical trial registration

NA.

## Study protocol

NA.

## Data collection

NA.

## Outcomes

NA.

# Sensorimotor experience remaps visual input to a heading-direction network

<https://doi.org/10.1038/s41586-019-1772-4>

Yvette E. Fisher<sup>1</sup>, Jenny Lu<sup>1</sup>, Isabel D'Alessandro<sup>1</sup> & Rachel I. Wilson<sup>1\*</sup>

Received: 28 December 2018

Accepted: 24 October 2019

Published online: 20 November 2019

In the *Drosophila* brain, ‘compass’ neurons track the orientation of the body and head (the fly’s heading) during navigation<sup>1,2</sup>. In the absence of visual cues, the compass neuron network estimates heading by integrating self-movement signals over time<sup>3,4</sup>. When a visual cue is present, the estimate of the network is more accurate<sup>1,3</sup>. Visual inputs to compass neurons are thought to originate from inhibitory neurons called R neurons (also known as ring neurons); the receptive fields of R neurons tile visual space<sup>5</sup>. The axon of each R neuron overlaps with the dendrites of every compass neuron<sup>6</sup>, raising the question of how visual cues are integrated into the compass. Here, using in vivo whole-cell recordings, we show that a visual cue can evoke synaptic inhibition in compass neurons and that R neurons mediate this inhibition. Each compass neuron is inhibited only by specific visual cue positions, indicating that many potential connections from R neurons onto compass neurons are actually weak or silent. We also show that the pattern of visually evoked inhibition can reorganize over minutes as the fly explores an altered virtual-reality environment. Using ensemble calcium imaging, we demonstrate that this reorganization causes persistent changes in the compass coordinate frame. Taken together, our data suggest a model in which correlated pre- and postsynaptic activity triggers associative long-term synaptic depression of visually evoked inhibition in compass neurons. Our findings provide evidence for the theoretical proposal that associative plasticity of sensory inputs, when combined with attractor dynamics, can reconcile self-movement information with changing external cues to generate a coherent sense of direction<sup>7–12</sup>.

The compass neurons in the *Drosophila* brain exhibit some resemblance to the head-direction cells of the mammalian brain<sup>13–16</sup>. Visual cues stabilize the tuning preferences of mammalian head-direction cells<sup>15</sup>, and when a visual cue is rotated to a new horizontal position, the preferences of all of the head-direction neurons rotate together<sup>14,16</sup>. It has been proposed that the mammalian head-direction system represents a ring attractor—a network in which global dynamics exhibit multiple stable states that unfold in a repeated sequence in response to an input<sup>7,17,18</sup>. However, we do not know how visual cues anchor the mammalian head-direction system at a mechanistic level. It has been suggested that Hebbian synaptic plasticity of visual inputs enforces the correct mapping between sensory cues and attractor network states<sup>7</sup>.

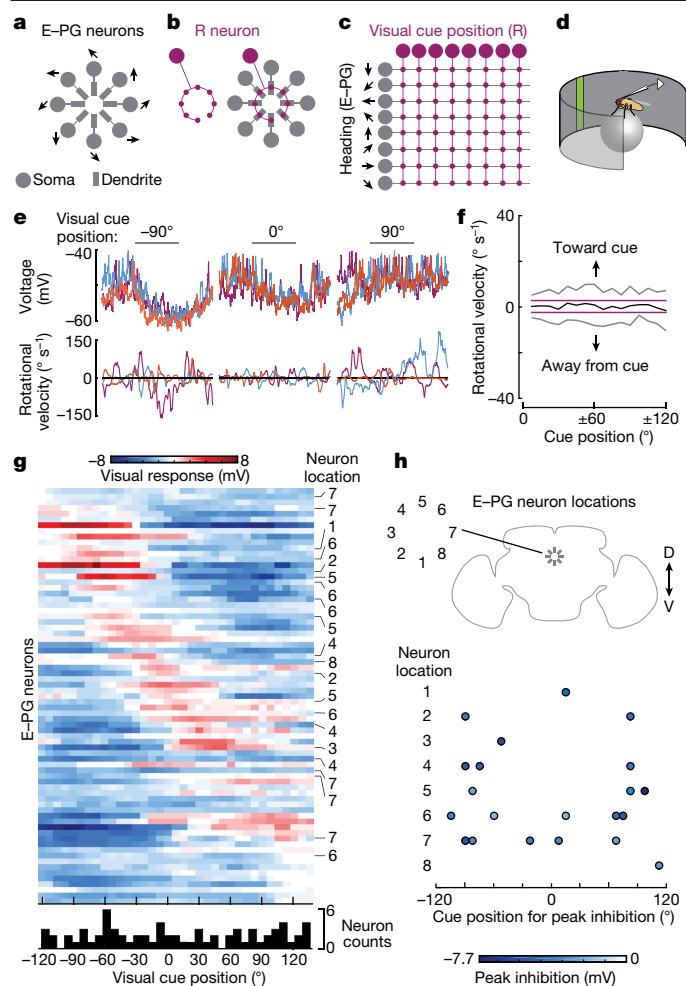
Similar to mammalian head-direction cells, *Drosophila* compass neurons (called E–PG neurons) have properties of a ring attractor<sup>2</sup>. Indeed, the dendrites of E–PG neurons are arranged in a ring in the brain (Fig. 1a). At any point in time, there is one ‘bump’ of activity in the E–PG ensemble, which rotates as the fly turns<sup>1</sup>. This network receives continuous input from brain regions that track the rotational velocity of the fly via optic flow signals, proprioceptive signals and/or motor efference signals<sup>3,4</sup>. These rotational velocity inputs push the bump around the circle. Visual cues make the position of the bump more accurate and stable<sup>1,3</sup>. We do not know whether visual inputs to E–PG

neurons are plastic: the offset between the E–PG bump and the visual world is different in different individuals and it can occasionally change unpredictably within an individual<sup>1,3</sup>; however, network instability alone does not provide evidence for synaptic plasticity.

The anatomy of R neuron axons is another reason to suspect the existence of synaptic plasticity in this network. Each R neuron axon overlaps with the dendrites of every E–PG neuron (Fig. 1b). If all these R-to-E–PG connections were functionally equivalent, information about the position of a visual cue would be discarded. Instead, it seems more likely that the all-to-all matrix of R-to-E–PG anatomical connections (Fig. 1c) represents a set of potential functional connections that can be repatterned during spatial learning. We therefore set out to test two hypotheses—first, that individual E–PG neurons respond selectively to specific visual cue positions and, second, that changes in visual-heading associations can trigger systematic, time-locked changes in the pattern of E–PG visual inputs.

Our first challenge was to isolate the synaptic input to E–PG neurons that is related to visual cue position, separate from the synaptic input related to the rotational velocity of the fly. We reasoned that this should be possible if we flashed visual cues transiently at randomized positions, preventing the fly from behaviourally fixating the stimulus. We therefore performed in vivo whole-cell recordings from E–PG

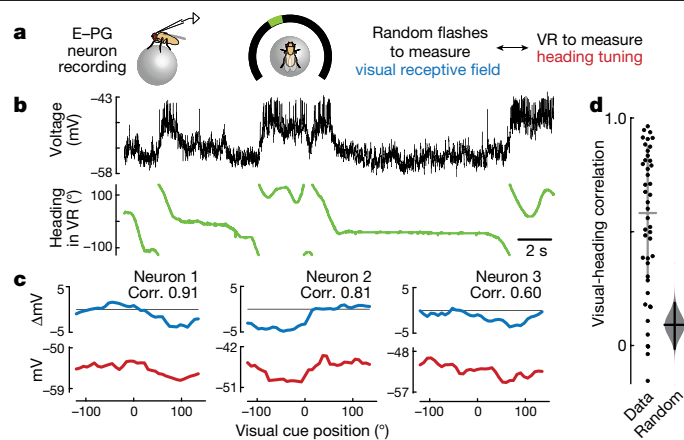
<sup>1</sup>Department of Neurobiology, Harvard Medical School, Boston, MA, USA. \*e-mail: [rachel\\_wilson@hms.harvard.edu](mailto:rachel_wilson@hms.harvard.edu)



**Fig. 1 | E-PG neurons are inhibited by visual cues at specific positions.**

**a**, E-PG neuron dendrites form a circular array, with adjacent cells representing adjacent headings. **b**, The axon of each R neuron forms a ring (left) that overlaps all E-PG dendrites (right). Magenta boutons represent presynaptic terminals. **c**, An unwrapped R-to-E-PG matrix. The receptive fields of R neurons tile visual space. **d**, An E-PG neuron is recorded in whole-cell mode while the fly walks on a ball, surrounded by a panorama in which a cue flashes at random horizontal positions. **e**, Top, three example E-PG responses per cue position, for three different positions, all from the same recording. Bottom, the rotational velocity of the same fly (+ indicates right; - indicates left). Note that the fly behaves differently on different trials, but the neural response is essentially the same regardless of the fly's behaviour. Cue flash is 500 ms. **f**, Mean rotational velocity around cue presentation (black),  $\pm 1$  s.d. (grey) across flies. Magenta lines show bootstrapped 95% confidence interval of the mean across flies after randomizing cue positions, Bonferroni-corrected; because the mean lies within these bounds, it is not significantly different from random. **g**, Summary of visual receptive fields of E-PG neurons (73 neurons in 68 flies). Cells are sorted by the cue position that evoked the most positive (least negative) response. Histogram shows the number of cells preferring each cue position. Some cells were filled to determine their location. **h**, Cue position eliciting peak inhibition versus neuron location (no significant correlation: circular correlation coefficient<sup>36</sup> = 0.097,  $P = 0.66$ ,  $n = 21$ ; see Extended Data Fig. 2g). D, dorsal; V, ventral.

neurons while flashing a bright vertical bar on a dark circular panorama at randomized horizontal positions (Fig. 1d). In a typical neuron, we observed hyperpolarization that was time-locked to flashes at specific positions (Fig. 1e). To verify that these neural responses are not related to the rotational velocity of the fly, we analysed the movement of the air-cushioned ball that the fly was standing on (Extended Data Fig. 1). Neural responses were unrelated to the rotational velocity of the flies



**Fig. 2 | Visual receptive fields of E-PG neurons align with heading tuning.**

**a**, Interleaved blocks measuring visual receptive fields and heading tuning. **b**, Top, E-PG voltage during a VR epoch. Bottom, VR heading. A heading of  $0^\circ$  means the cue is in front of the fly. **c**, Comparison of visual receptive fields (blue) and heading tuning (red) from three example E-PG neurons (from three flies, with Pearson's correlation coefficients). **d**, Pearson's correlation coefficients from 40 cells in 39 flies (all cells from Fig. 1). The mean and 95% confidence interval are shown as horizontal and vertical lines, respectively. The means of the data are outside the 95% confidence interval of a bootstrap distribution (grey violin plot) computed on randomized visual-heading pairings.

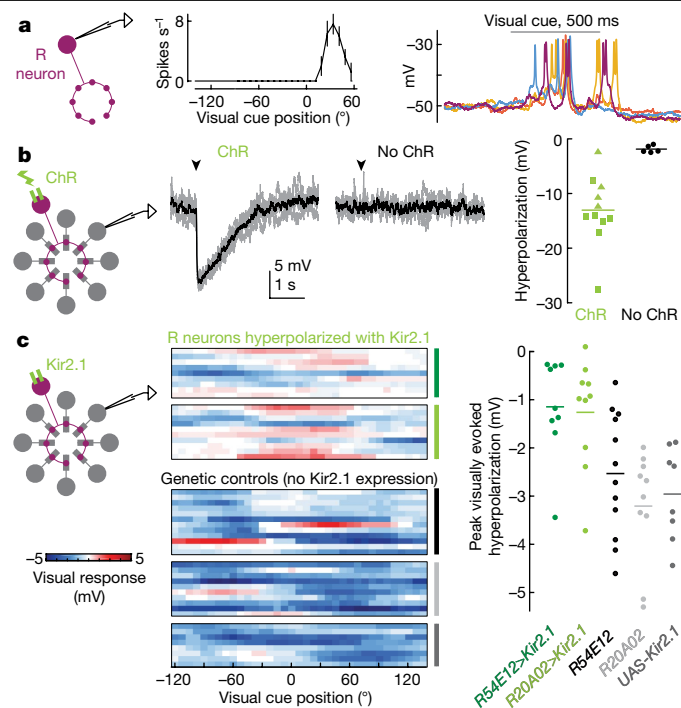
around the time of the visual flash (Fig. 1e), and there was no correlation between the rotational velocity and the flash (Fig. 1f). Therefore, we can interpret visually locked responses as synaptic inputs related to visual cue position. We call this the visual receptive field of the cell. The finding of visually evoked hyperpolarization is consistent with the fact that R neurons release the inhibitory neurotransmitter  $\gamma$ -aminobutyric acid (GABA)<sup>19,20</sup>.

In almost every E-PG neuron, we found that some visual cue positions elicited hyperpolarization while other positions elicited no hyperpolarization (Fig. 1e, g). This suggests that each E-PG neuron receives relatively strong input from some R neurons but weak or non-existent input from other R neurons. In approximately half of the E-PG neurons, we also found that some cue positions elicited depolarization (Fig. 1g). Depolarization may represent disinhibition: because there is ongoing mutual inhibition between E-PG neurons<sup>2</sup>, a visual cue that inhibits one E-PG neuron will disinhibit other E-PG neurons.

We found that different E-PG neurons had distinct visual receptive fields (Fig. 1g). When we sorted cells by the position that elicited the most positive (least negative) response, we found a uniform mapping of cue positions onto E-PG neurons (Fig. 1g). However, cue-evoked hyperpolarization was more prominent for lateral cue positions (Extended Data Fig. 2); this spatial bias is probably inherited from R neurons, because the receptive fields of R neurons are similarly biased towards lateral positions<sup>5</sup>.

When we managed to record sequentially from two adjacent E-PG neurons in the same brain, we found that they had adjacent receptive fields, as expected (Extended Data Fig. 3). However, when we pooled data across brains, we found no systematic relationship between the location of the dendrites of the E-PG neuron and its receptive field (Fig. 1h). Therefore, the mapping from visual space to compass coordinates is different across individuals.

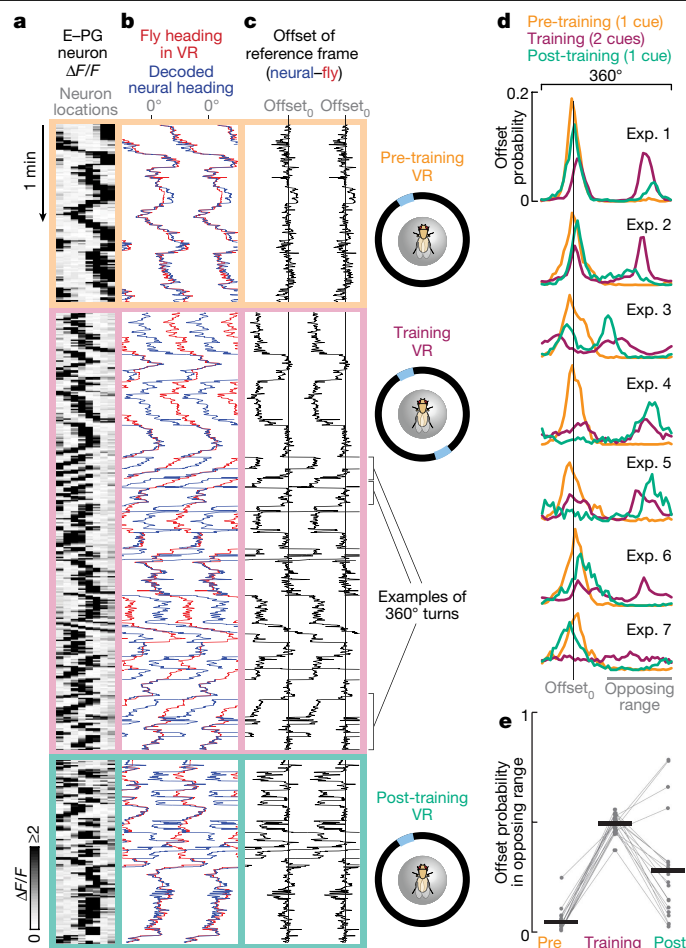
Next, we investigated how the visual receptive field of an E-PG neuron compares with its heading tuning. To measure heading tuning, we allowed the fly to walk in closed-loop virtual reality (VR) in which the horizontal position of the cue was locked to the virtual heading of the fly (Fig. 2a, b). We periodically paused VR to map the visual receptive



**Fig. 3 | R neurons drive visually evoked inhibition in E-PG neurons.** **a**, Left, visually evoked spike rates in an R neuron (mean  $\pm$  s.e.m. across trials,  $n = 5-6$  trials, R2 neurons). Right, four responses to repeated presentation of the best cue position for this neuron. We observed spatially tuned responses in 3 out of 7 R2 cells and 1 out of 3 R4d cells; an additional 3 R2 cells and 1 R4d cell responded to full-field illumination but were unresponsive to the cue or not spatially tuned. **b**, Left, responses of an E-PG neuron to optogenetic activation of R2 neurons via Chrimson (ChR), with four single trials in grey, mean in black. Middle, same as for the left panel, but with no Chrimson in R neurons,  $n = 4$  trials. Right, summary of mean evoked hyperpolarization with Chrimson in R neurons (squares, R2 neurons,  $n = 7$ ; triangles, R4d neurons,  $n = 4$ ) and controls ( $n = 5$ ). **c**, Left, E-PG visual receptive fields in flies in which R neurons were chronically hyperpolarized using Kir2.1 expression driven by *R54E12-Gal4* or *R20A02-Gal4* (green shades) versus controls (*R54E12-Gal4* only, *R20A02-Gal4* only, *UAS-Kir2.1* only, grey shades). Right, summary of peak visually evoked hyperpolarization, colour-coded as in the left panel (horizontal lines are means;  $n = 9, 10, 12, 10, 8$  cells, from left to right; *R54E12-Kir2.1* versus *R54E12/+* and *UAS/+*,  $P = 0.021$  and  $P = 0.0016$ , respectively; *R20A02-Kir2.1* versus *R20A02/+* and *UAS/+*,  $P = 0.0046$  and  $P = 0.012$ , respectively; two-sided Wilcoxon rank-sum tests).

field of the same neuron using brief random flashes. In most neurons, we found that the visual receptive field was correlated with heading tuning (Fig. 2c, d and Extended Data Figs. 3, 4). This result is notable because heading tuning reflects not only synaptic inputs related to visual cue position, but also synaptic inputs related to the rotational velocity of a fly. Imperfect alignment between these inputs may explain why some neurons showed poor correlations (Fig. 2d).

To confirm that R neurons are the actual source of visual responses in E-PG cells, we focused on two R neuron types (R2 and R4d) that respond to sparse visual cues<sup>5</sup>. First, we used whole-cell recordings to confirm that these R neuron types can be excited by the visual cue (Fig. 3a). Second, we verified that optogenetically activating either R2 or R4d neurons inhibits E-PG neurons (Fig. 3b). Third, we established that R neurons are required for normal visually evoked hyperpolarization in E-PG neurons. We used two independent driver lines to hyperpolarize R2 or R4d neurons by overexpressing the potassium channel Kir2.1 (Extended Data Fig. 5), and we confirmed that visually evoked hyperpolarization was attenuated (Fig. 3c and Extended Data Fig. 6) in both

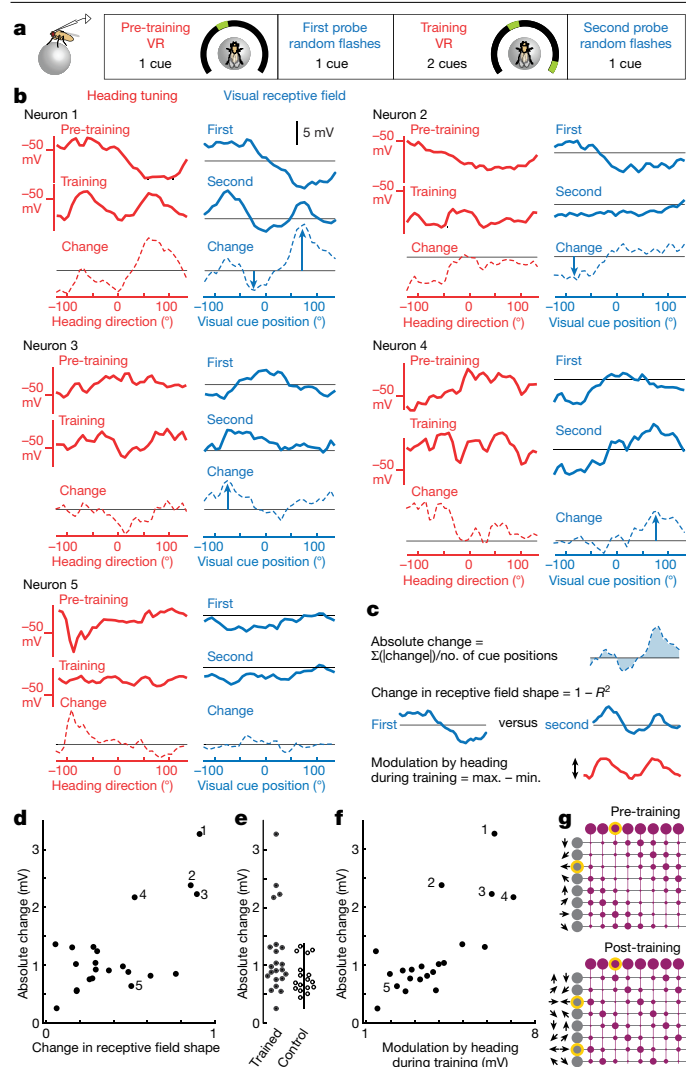


**Fig. 4 | Visuomotor experience can persistently change E-PG ensemble representations of heading direction.** **a**, E-PG ensemble GCaMP6f signals. Here the circular E-PG ensemble has been linearized, with each row showing eight sectors of the ensemble. The fly walked in a one-cue environment (pre-training,  $\geq 10$  min), then a two-cue environment (training, 20 min), and finally a one-cue environment (post-training,  $\geq 4$  min). Three snippets of one experiment are shown. Brackets mark 360° turns when the bump skipped over half the ensemble. **b**, In the same experiment, VR heading (red) overlaid with the decoded neural representation of heading (blue). We double-plotted both traces and shifted the entire red trace horizontally so that it overlapped with the blue trace during pre-training. **c**, The offset of the decoded neural representation of heading relative to VR heading, double-plotted. The circular mean during pre-training is marked with a vertical line (defined as  $\text{offset}_0$ ). **d**, Offset probability histograms during each block, for seven example experiments. We found diverse values of  $\text{offset}_0$  in different flies, as reported previously<sup>1</sup>, but for display we horizontally aligned all  $\text{offset}_0$  values in different flies. The opposing range is defined as the range from ( $\text{offset}_0 + 90^\circ$ ) to ( $\text{offset}_0 - 90^\circ$ ). **e**, Total offset probability in the opposing range. Each set of connected points is one experiment ( $n = 19$  flies). Training and post-training are both significantly different from pre-training ( $P = 3.8 \times 10^{-6}$  training versus pre-training and  $P = 5.31 \times 10^{-5}$  post-training versus pre-training, two-sided exact paired Wilcoxon signed-rank tests).

genotypes. A few E-PG neurons still showed some visual responses, probably because neither driver line achieved complete coverage of R2 and R4d neurons (Extended Data Figs. 5, 6).

Next, we turned to our second hypothesis—that changes in visual-heading associations can trigger systematic, time-locked changes in the visual receptive fields of E-PG neurons. After allowing the fly to navigate in VR with one visual cue (the pre-training block), we switched to VR with two cues positioned 180° apart (the training block). In the training block, a full turn and a half-turn will arrive at an identical view of





**Fig. 5 | Visuomotor experience can remap visual input to E-PG neurons contingent on postsynaptic activity.** **a**, After the fly navigated in VR with one cue (pre-training), we measured the visual receptive field of an E-PG neuron (first probe). Then the fly navigated in VR with two cues for 12 min (training) and we again measured the visual receptive field (second probe). **b**, Five example neurons. For each neuron, the red solid curve is heading tuning. Red dashed curve is the change in heading tuning (training minus pre-training). Blue solid curve is the visual receptive field. Blue dashed curve is the change in the visual receptive field (second probe minus first probe). Arrowheads mark large changes. Black vertical scale bar applies to all data in **b**. Thin black horizontal lines indicate zero change ( $\Delta = 0$  mV). Neuron 5 is an example with little modulation by heading during training and little change in visual receptive field. **c**, Explanation of metrics in **d–f**. **d**, Absolute change in visual receptive field, versus change in receptive field shape ( $R^2 = 0.44$ ,  $P = 0.00078$  testing  $t$ -statistic slope  $\neq 0$ ; 22 E-PG neurons in 22 flies). **e**, Absolute change in visual receptive field post-training (22 E-PG neurons in 22 flies) versus controls (17 E-PG neurons in 17 flies). Post-training flies are significantly different from control flies ( $P = 0.043$ , two-sided Wilcoxon rank-sum test). Controls walked in a one-cue VR (not two-cue VR) between the first and second probe. Four training experiments had changes significantly larger than any controls ( $>2$  s.d. above control mean, vertical bar); these are neurons 1–4. **f**, Absolute change in visual receptive field, versus modulation by heading during training ( $R^2 = 0.52$ ,  $P = 0.00016$  testing  $t$ -statistic slope  $\neq 0$ ; 22 E-PG neurons in 22 flies). **g**, Schematic of model. When a visual cue appears, it activates specific R neurons (highlighted magenta cell) and this pushes the bump towards the E-PG neuron with minimal inhibition (highlighted grey cell). Training changes R-to-E-PG weights so that the bump toggles between two offsets during post-training. R neurons are ordered by receptive field position, and E-PG neurons are ordered by preferred heading direction (arrows).

the world, meaning the correlation between rotational velocity signals and visual cue position signals will be altered.

To assess the effect of training on network dynamics, we imaged calcium signals from the entire E-PG ensemble (Fig. 4a). During pre-training, there was a stable offset between the visual environment and the E-PG bump (Fig. 4b, c). During training, the offset toggled between two values approximately  $180^\circ$  apart. This result is expected, because there are two equally valid interpretations of the visual scene, yet only one bump can exist in the E-PG ensemble<sup>2</sup>. When the fly made a  $360^\circ$  turn, we often saw the bump flow twice around  $180^\circ$  of the E-PG ensemble, skipping over the other  $180^\circ$  (Fig. 4a–c). Rotational velocity inputs to the E-PG network should drive the bump to traverse the full circle during a full turn<sup>3,4</sup>; the skipping-over phenomenon thus indicates the dominance of visual position inputs over angular velocity inputs. The E-PG neurons that were traversed twice essentially displayed two preferred heading directions; this is reminiscent of the finding that some rat head-direction cells show two preferred directions in an environment with twofold rotational symmetry<sup>21</sup>.

Upon returning to a one-cue environment (post-training), the offset sometimes immediately settled into its original value. Often, however, this was not the case. Rather, the offset continued to toggle for several minutes, or else it immediately settled in a new value rather than the original one (Fig. 4d, e). Both of the latter two outcomes suggest a persistent, systematic change in the way that visual cues are mapped onto E-PG neurons. We observed one of the latter outcomes in about half of our experiments (Fig. 4e and Extended Data Fig. 7).

Finally, to investigate whether training changes visual receptive fields, we returned to E-PG whole-cell recordings (Fig. 5a). We began each experiment with one visual cue in VR (pre-training). We then switched to two visual cues in VR (training). Between each block of VR, we periodically paused to map the receptive field of the neuron using brief random flashes. Whereas we used a  $360^\circ$  panorama during calcium imaging, the spatial constraints of electrophysiology required us to map the  $360^\circ$  environment onto a  $270^\circ$  panorama<sup>1,10</sup>.

During the training block, we found that some E-PG neurons were strongly modulated by the fly's heading. In these neurons, training produced changes in the visual receptive field (Fig. 5b, neurons 1–4). These changes were bidirectional, suggesting that visually evoked inhibition was depressed for some cue locations and potentiated for others. We quantified these changes by summing the absolute value of the change in the receptive field across all cue positions (absolute change; Fig. 5c). We also measured the change in the shape of the receptive field (Fig. 5c). These metrics were correlated across experiments (Fig. 5d); we never saw a large absolute change in the receptive field without a change in receptive field shape. We also never observed large changes in the receptive field under control conditions in which flies only experienced one cue in VR (but not two cues) during the period between the receptive field mapping epochs (Fig. 5e, Extended Data Figs. 8, 9).

By contrast, other E-PG neurons were essentially unmodulated by the fly's heading during training (Fig. 5b, neuron 5). These neurons may reside in sectors of the ensemble that were skipped over by the bump during training. Notably, training had almost no effect on visual receptive fields in these neurons (Fig. 5b and Extended Data Fig. 8). Overall, the magnitude of heading modulation during training was significantly correlated with the subsequent change in the visual receptive field (Fig. 5f). This correlation indicates that remapping depends on the activity of the E-PG neuron. Simply exposing the fly to the altered visual environment is not sufficient; rather, visual cues must intersect with heading representations in E-PG neurons. Because R-to-E-PG synapses are the site of intersection between visual responses and heading representations, they are the most likely locus of plasticity. In a companion study, Kim et al.<sup>22</sup> used optogenetic manipulations to reach the same conclusion. Because R neuron dendrites form a retinotopic map that is fairly consistent across flies<sup>5</sup>, it seems unlikely that the visual map in R neuron dendrites is experience-dependent, further supporting the notion that R-to-E-PG synapses are the locus of plasticity.

## Discussion

We propose that correlated pre- and postsynaptic activity triggers associative long-term synaptic depression of R-to-E-PG inhibition. This learning rule would explain why visual receptive fields and heading tuning are typically aligned in E-PG neurons. When an individual R neuron is activated by a visual cue, it should push the bump of activity towards the E-PG neurons that it inhibits most weakly (Fig. 5g). If the full ring attractor network agrees with this outcome, then long-term synaptic depression will occur and those weak R-to-E-PG synapses will become even weaker, further reinforcing this outcome. To ensure network stability, long-term synaptic depression should be balanced by long-term potentiation at R-to-E-PG synapses; the co-existence of long-term synaptic depression and long-term potentiation would also explain why we found bidirectional changes in visual receptive fields after training (Fig. 5b). These learning rules should produce a doubled pattern of R-to-E-PG synaptic weights after training in a two-cue world (Fig. 5g), reflecting the twofold symmetry of visuomotor correlations.

The key result of this study—that visual inputs to E-PG neurons are plastic—supports theoretical models that describe how a network can progressively establish a spatial map of the world by incorporating information about consistent sensory cues during exploration<sup>7–12</sup>. In robotics, this process is called simultaneous localization and mapping<sup>23</sup>. Our results provide direct experimental evidence for this type of unsupervised learning at the level of synaptic potentials in vivo.

In a simultaneous localization and mapping framework, visual cues are often local, meaning that they can change in size and apparent angle as they are approached; by contrast, we chose to use visual cues that could not be approached, simplifying the relationship between heading and visual cues. This choice was motivated by the known receptive field properties of R2 or R4d neurons, which seem adapted to detect the position of the Sun (or Moon). Specifically, R2 or R4d neurons have large inhibitory surrounds, meaning that they only respond robustly to isolated visual objects<sup>5,24</sup> such as the Sun. The Sun is an ideal compass cue because it is effectively at infinity<sup>25</sup>.

We propose that plasticity at R-to-E-PG synapses allows the position of the Sun to be flexibly associated with other compass cues, such as the pattern of linearly polarized light in the sky<sup>26</sup>, sky-wide chromatic and intensity gradients<sup>27,28</sup>, and wind<sup>29,30</sup>. In other insects, the E-PG network responds to multiple sorts of compass cues<sup>31,32</sup>, and navigation behaviour can depend on arbitrary learned associations between compass cues<sup>33–35</sup>. In a companion study, Kim et al.<sup>22</sup> provide evidence in favour of the idea that plasticity could be used to learn a complex conjunction of visual objects; in the future, to test this idea, it will be interesting to see whether any complex scene can generate a progressively more-stable heading representation (offset) during training. It will also be important to extend the approach that we have taken here to simulate a more naturalistic virtual world, to study how multiple types of cues influence the behaviour of this network and the organism.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1772-4>.

1. Seelig, J. D. & Jayaraman, V. Neural dynamics for landmark orientation and angular path integration. *Nature* **521**, 186–191 (2015).
2. Kim, S. S., Rouault, H., Druckmann, S. & Jayaraman, V. Ring attractor dynamics in the *Drosophila* central brain. *Science* **356**, 849–853 (2017).

3. Green, J. et al. A neural circuit architecture for angular integration in *Drosophila*. *Nature* **546**, 101–106 (2017).
4. Turner-Evans, D. et al. Angular velocity integration in a fly heading circuit. *eLife* **6**, e23496 (2017).
5. Seelig, J. D. & Jayaraman, V. Feature detection and orientation tuning in the *Drosophila* central complex. *Nature* **503**, 262–266 (2013).
6. Omoto, J. J. et al. Neuronal constituents and putative interactions within the *Drosophila* ellipsoid body neuropil. *Front. Neural Circuits* **12**, 103 (2018).
7. Skaggs, W. E., Knierim, J. J., Kudrimoti, H. S. & McNaughton, B. L. A model of the neural basis of the rat's sense of direction. *Adv. Neural Inf. Process. Syst.* **7**, 173–180 (1995).
8. Milford, M. J., Wyeth, G. F. & Prasser, D. RatSLAM: A hippocampal model for simultaneous localization and mapping. In *Proc. International Conference on Robotics and Automation* 403–408 (2004).
9. Mulas, M., Waniek, N. & Conradt, J. Hebbian plasticity realigns grid cell activity with external sensory cues in continuous attractor models. *Front. Comput. Neurosci.* **10**, 13 (2016).
10. Cope, A. J., Sabo, C., Vasilaki, E., Barron, A. B. & Marshall, J. A. A computational model of the integration of landmarks and motion in the insect central complex. *PLoS ONE* **12**, e0172325 (2017).
11. Keinath, A. T., Epstein, R. A. & Balasubramanian, V. Environmental deformations dynamically shift the grid cell spatial metric. *eLife* **7**, e38169 (2018).
12. Ocko, S. A., Hardcastle, K., Giocomo, L. M. & Ganguli, S. Emergent elasticity in the neural code for space. *Proc. Natl Acad. Sci. USA* **115**, E11798–E11806 (2018).
13. Taube, J. S., Muller, R. U. & Ranck, J. B. Jr. Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. *J. Neurosci.* **10**, 420–435 (1990).
14. Taube, J. S., Muller, R. U. & Ranck, J. B. Jr. Head-direction cells recorded from the postsubiculum in freely moving rats. II. Effects of environmental manipulations. *J. Neurosci.* **10**, 436–447 (1990).
15. Mizumori, S. J. & Williams, J. D. Directionally selective mnemonic properties of neurons in the lateral dorsal nucleus of the thalamus of rats. *J. Neurosci.* **13**, 4015–4028 (1993).
16. Knierim, J. J., Kudrimoti, H. S. & McNaughton, B. L. Place cells, head direction cells, and the learning of landmark stability. *J. Neurosci.* **15**, 1648–1659 (1995).
17. Zhang, K. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *J. Neurosci.* **16**, 2112–2126 (1996).
18. Xie, X., Hahnloser, R. H. & Seung, H. S. Double-ring network model of the head-direction system. *Phys. Rev. E* **66**, 041902 (2002).
19. Hanesch, U., Fischbach, K. F. & Heisenberg, M. Neuronal architecture of the central complex in *Drosophila melanogaster*. *Cell Tissue Res.* **257**, 343–366 (1989).
20. Zhang, Z., Li, X., Guo, J., Li, Y. & Guo, A. Two clusters of GABAergic ellipsoid body neurons modulate olfactory labile memory in *Drosophila*. *J. Neurosci.* **33**, 5175–5181 (2013).
21. Jacob, P. Y. et al. An independent, landmark-dominated head-direction signal in dysgranular retrosplenial cortex. *Nat. Neurosci.* **20**, 173–175 (2017).
22. Kim, S. S., Hermundstad, A. M., Romani, S., Abbott, L. F. & Jayaraman, V. Generation of stable heading representations in diverse visual scenes. *Nature* <https://doi.org/10.1038/s41586-019-1767-1> (2019).
23. Cadena, C. et al. Past, present, and future of simultaneous localization and mapping: toward the robust-perception age. *IEEE Trans. Robot.* **32**, 1309–1332 (2016).
24. Sun, Y. et al. Neural signatures of dynamic stimulus selection in *Drosophila*. *Nat. Neurosci.* **20**, 1104–1113 (2017).
25. Wehner, R. Astronavigation in insects. *Annu. Rev. Entomol.* **29**, 277–298 (1984).
26. Wehner, R. & Müller, M. The significance of direct sunlight and polarized skylight in the ant's celestial system of navigation. *Proc. Natl Acad. Sci. USA* **103**, 12575–12579 (2006).
27. el Jundi, B., Smolka, J., Baird, E., Byrne, M. J. & Dacke, M. Diurnal dung beetles use the intensity gradient and the polarization pattern of the sky for orientation. *J. Exp. Biol.* **217**, 2422–2429 (2014).
28. el Jundi, B., Foster, J. J., Byrne, M. J., Baird, E. & Dacke, M. Spectral information as an orientation cue in dung beetles. *Biol. Lett.* **11**, 20150656 (2015).
29. Bell, W. J., Tobin, T. R. & Sorensen, K. A. Orientation responses of individual larder beetles, *Dermestes ater* (Coleoptera, Dermestidae), to directional shifts in wind stimuli. *J. Insect Behav.* **2**, 787–801 (1989).
30. Heinzel, H.-G. & Böhm, H. The wind-orientation of walking carrion beetles. *J. Comp. Physiol. A* **164**, 775–786 (1989).
31. el Jundi, B. et al. Neural coding underlying the cue preference for celestial orientation. *Proc. Natl Acad. Sci. USA* **112**, 11395–11400 (2015).
32. Pegel, U., Pfeiffer, K. & Homberg, U. Integration of celestial compass cues in the central complex of the locust brain. *J. Exp. Biol.* **221**, jeb171207 (2018).
33. Müller, M. & Wehner, R. Wind and sky as compass cues in desert ant navigation. *Naturwissenschaften* **94**, 589–594 (2007).
34. el Jundi, B. et al. A snapshot-based mechanism for celestial orientation. *Curr. Biol.* **26**, 1456–1462 (2016).
35. Dacke, M. et al. Multimodal cue integration in the dung beetle compass. *Proc. Natl Acad. Sci. USA* **116**, 14248–14253 (2019).
36. Jammalamadaka, S. R. & SenGupta, A. *Topics in Circular Statistics* (World Scientific, 2001).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

## Methods

### Fly husbandry and genotypes

Unless otherwise stated, flies were raised on standard cornmeal-molasses food (New Brown 19L, Archon Scientific) in an incubator on a 12-h:12-h light:dark cycle at 25 °C with humidity between around 50 and 70%.

All experiments with visual stimuli used flies with at least one wild-type copy of the *white* gene, and most electrophysiology experiments used flies with two copies of the wild-type *white* gene (as described below).

The experimenter was not blind to genotype because we did not use genetic perturbations; the exception is Fig. 3c (Kir2.1 perturbation). For the dataset shown in Fig. 3c, the experimenter was blind to genotype after the pilot phase for driver line *R20A02-Gal4*; because Fig. 3c pilot data were indistinguishable from subsequent data, all data were ultimately pooled; overall the experimenter was blind to genotype in 67% of these recordings. For the dataset obtained using the driver line *R54E12-Gal4*, the experimenter was not blind to genotype because the experimental genotype was obtained at a lower-than-expected (sub-Mendelian) frequency, making it impractical to blind the experimenter.

Genotypes of fly stocks used in each figure are as follows. For Figs. 1, 2, 5 and Extended Data Figs. 1–4, 8, 9, we used *P{20XUAS-IVS-mCD8::GFP}attP40/P{20XUAS-IVS-mCD8::GFP}attP40;P{R60D05-Gal4}attP2/P{R60D05-Gal4}attP2* flies. For Fig. 3a, we used *P{20XUAS-IVS-mCD8::GFP}attP40/P{20XUAS-IVS-mCD8::GFP}attP40;P{GawB}EB1/+* flies. For Fig. 3b (R2 activation), we used *w/+;P{R19C08-lexA}attP40/P{20XUAS-IVS-mCD8::GFP}attP40;PBac{13xLexAop2-IVS-Syn21-Chrimson::tdT-3.1}VK00005/P{R60D05-Gal4}attP2* flies. For Fig. 3b (R4d activation), we used *w/+;P{R60D05-lexA}attP40/P{13XLexAop2-mCD8::GFP}attP40;P{20XUAS-CsChrimson-tdTomato}VK00005/P{R12B01-Gal4}attP2* flies. For Fig. 3b (no Chrimson), we used *P{20XUAS-IVS-mCD8::GFP}attP40/P{20XUAS-IVS-mCD8::GFP}attP40;P{R60D05-Gal4}attP2/P{R60D05-Gal4}attP2* flies. For Fig. 3c and Extended Data Fig. 6 (Kir2.1 silencing, driver 1), we used *+w;P{R60D05-lexA}attP40/P{13XLexAop2-mCD8::GFP}attP40;P{R20A02-Gal4}attP2/P{UAS-Hsap|KCNJ2.eGFP}3* flies. For Fig. 3c and Extended Data Fig. 6 (Kir2.1 silencing, driver 2), we used *+w;P{R60D05-lexA}attP40/P{13XLexAop2-mCD8::GFP}attP40;P{R54E12-Gal4}attP2/P{UAS-Hsap|KCNJ2.eGFP}* flies. For Fig. 3c and Extended Data Fig. 6 (UAS-only controls), we used *+w;P{R60D05-lexA}attP40/P{13XLexAop2-mCD8::GFP}attP40;+/P{UAS-Hsap|KCNJ2.eGFP}3* flies. For Fig. 3c and Extended Data Fig. 6 (Gal4-only controls), we used *+w;P{GMR60D05-lexA}attP40/P{13XLexAop2-mCD8::GFP}attP40;+/P{R20A02-Gal4}attP2* flies and *+w;P{GMR60D05-lexA}attP40/P{13XLexAop2-mCD8::GFP}attP40;+/P{R54E12-Gal4}attP2* flies. For Fig. 4 and Extended Data Fig. 7, we used *+w;P{UAS-GCaMP6f}attP40/+;P{R60D05-Gal4}attP2/+* flies. For Extended Data Fig. 5, we used *R57C10-FLPG5.PEST;UAS(FRT.stop)myr::smGdP-HA, UAS(FRT.stop)myr::smGdP-V5, UAS(FRT.stop)myr::smGdP-Flag/R20A02-Gal4, R57C10-FLPG5.PEST;UAS(FRT.stop)myr::smGdP-HA, UAS(FRT.stop)myr::smGdP-V5, UAS(FRT.stop)myr::smGdP-Flag/R54E12-Gal4* flies.

### Origins of transgenic stocks

The following GMR Gal4 lines were obtained from the Bloomington *Drosophila* Stock Center (BDSC) and were described previously<sup>37</sup>: *P{R60D05-Gal4}attP2*, *P{R60D05-lexA}attP40*, *P{R19C08-lexA}attP40*, *P{R12B01-Gal4}attP2*, *P{R54E12-Gal4}attP2*, *P{R20A02-Gal4}attP2*. The *P{GawB}EB1* line was also obtained from the BDSC and was described previously<sup>38</sup>.

*P{20XUAS-IVS-mCD8::GFP}attP40* was a gift from B. Pfeiffer and G. Rubin and was described previously<sup>39</sup>. *P{13XLexAop2-mCD8::GFP}attP40* was obtained from the BDSC and was described previously<sup>39</sup>. *PBac{13xLexAop2-IVS-Syn21-Chrimson::tdT-3.1}VK00005* was a gift from B. Pfeiffer and D. Anderson and was described previously<sup>40</sup>. *P{20X-UAS-CsChrimson-tdTomato}VK00005* was a gift from J. Tuthill who obtained

it from B. Pfeiffer. (Note that we have confirmed that this CsChrimson insert is on the third chromosome, but it may not be in *VK00005*, given the recombination frequencies observed in our laboratory. We have confirmed that this insertion does generate tdTomato expression and light-evoked currents in Gal4<sup>+</sup> cells.) *P{UAS-Hsap|KCNJ2.eGFP}7* was obtained from the BDSC and was described previously<sup>41</sup>. *P{UAS-GCamp6f}attP40* was obtained from the BDSC through T. Clandinin and was described previously<sup>42</sup>.

Transgenes for MultiColor FlpOut were obtained from the BDSC and were described previously<sup>43</sup>, including *w[1118]P{y[+t7.7]w[+mC]}=GMR57C10-FLPG5.PEST}su(Hw)attP8;PBac{y[+mDint2]}* and *w[+mC]=10xUAS(FRT.stop)myr::smGdP-HA|VK00005P{y[+t7.7]}* and *w[+mC]=10xUAS(FRT.stop)myr::smGdP-V5-THS-10xUAS(FRT.stop)myr::smGdP-Flag}su(Hw)attP1*.

### Fly preparation and dissection

Newly eclosed virgin female flies were anaesthetized on ice (electrophysiology) or CO<sub>2</sub> (imaging) and were collected around 3–10 h (electrophysiology) or 12–26 h (imaging) before the experiment. In some cases, to promote walking behaviour, we deprived the flies of food (but not water) for approximately 3–10 h before the experiment, and experiments were performed around the subjective evening of the fly ( $\pm 2$  h from light to dark switch, Zeitgeber time 12); this was done in Fig. 5 and in 72% of recordings in Fig. 2. In all other experiments, there was not circadian restriction and flies were kept on food until the dissection. At the beginning of each dissection, the fly was cold-anaesthetized.

For electrophysiology experiments, the preparation holder consisted of flat titanium foil secured to an acrylic platform, with the foil oriented parallel to the horizontal body plane; the fly's head and body were gently pushed partway-through a hole in the foil. For E–PG neuron electrophysiology, the head was pitched forward so that the posterior surface was roughly parallel to the foil and most of each eye was under the foil. For R neuron electrophysiology, the head was positioned in a more upright angle, and a 90° bend was made in the foil to maximize the area of the eyes that was under the foil. For imaging experiments, the preparation holder was shaped like an inverted pyramid and was CNC machined from black acrylic (Autotiv), and the head was pitched forward so that the posterior surface was oriented dorsally and most of the eye was under the holder. The fly was always secured in the holder with epoxy (Loctite AA 3972) and cured using a brief (<1-s) pulse of ultraviolet light (LED-200, Electro-Lite Co). Wings were sometimes repositioned or removed. After the dorsal head was covered in saline, a hole was cut in the head capsule and some trachea were removed to expose the brain area of interest. To reduce brain movement, muscle 16 was removed, the proboscis was removed (Figs. 1–3, 5) or glued (Fig. 4) and the oesophagus was clipped or removed (Fig. 4). For electrophysiology, an aperture was made in the perineural sheath around the somata of interest either by ripping gently with fine forceps or by using suction from a patch pipette filled with external solution.

The external solution contained (in mM): 103 NaCl, 3 KCl, 5 *N*-tris(hydroxymethyl) methyl-2-aminoethane-sulfonic acid, 8 trehalose, 10 glucose, 26 NaHCO<sub>3</sub>, 1 NaH<sub>2</sub>PO<sub>4</sub>, 1.5 CaCl<sub>2</sub> and 4 MgCl<sub>2</sub>, with osmolarity adjusted to 270–273 mOsm. External solution was bubbled with 95% O<sub>2</sub> and 5% CO<sub>2</sub> and reached a final pH of 7.3. External solution was continuously perfused over the brain during electrophysiology and before imaging.

### Patch-clamp recordings

Patch pipettes were made from borosilicate glass (Sutter, 1.5 mm o.d., 86 mm i.d.) using a Sutter P-97 puller. For E–PG recordings, the pipette was fire-polished after pulling<sup>44</sup> using a microforge (ALA Scientific Instruments) to achieve a final resistance of 8–15 MΩ. For R neuron recordings, pipettes (4–10 MΩ) were not fire-polished. The internal solution contained (in mM): 140 potassium aspartate, 10 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid, 4 MgATP,

0.5 Na<sub>3</sub>GTP, 1 ethylene glycol tetraacetic acid, 1 KCl and 13 biocytin hydrazide. The pH was 7.3 and the osmolarity was adjusted to approximately 268 mOsm. To encourage walking, the external solution was heated before the experiment<sup>45</sup> to around 25–32 °C; this was done for all recordings in Fig. 5, 65% of recordings in Fig. 2 and 42% of recordings in Fig. 1. All other recordings were performed using external solution at room temperature.

To obtain patch-clamp recordings under visual control, we used an Olympus BX51WI microscope with a 40× water-immersion objective. Neurons were identified as GFP<sup>+</sup> using a Hg-lamp source (U-LH100HG, Olympus) with an eGFP long-pass filter (U-N41012, Chroma). For experiments in which the fly was positioned on a foam ball, farred light was delivered from a fibre-coupled LED (740 nm, M740F2, Thorlabs) through a ferrule patch cable (200 µm core, Thorlabs) plugged into a fibre-optic cannula (1.25 mm SS ferrule 200 µm core, 0.22 NA, Thorlabs) glued to the recording platform, with the tip of the cannula around 1 cm behind the fly. In experiments without the ball, the brain was illuminated with 780 nm light via the microscope condenser, and after the recording was obtained, the condenser was lowered to prevent it from obscuring the fly's view of the visual panorama.

Recordings were obtained using an Axopatch 200B amplifier and a CV-203BU headstage (Molecular Devices). Voltage signals were low-pass filtered at 5 kHz before digitalization and then acquired with a NiDAQ PCI-6251 (National Instruments) at 20 kHz. Liquid junction-potential correction was performed post hoc by subtracting 13 mV from recorded voltages<sup>46</sup>.

### Two-photon calcium imaging

Imaging experiments were performed using a two-photon microscope with a moveable stage (Thorlabs Bergamo II) and a fast piezoelectric objective scanner (Physik Instrument P725) for volumetric imaging. For two-photon excitation, we used a Chameleon Vision-S Ti:Sapphire femtosecond laser tuned to 940 nm. Images were collected using a 20×/1.0 NA objective (Olympus). Emission fluorescence was filtered with a 525-nm bandpass filter (Thorlabs) and collected using a GASP photomultiplier tube (Hamamatsu).

The imaging region was centred on the protocerebral bridge, which is the brain region in which E–PG neuron axons terminate. For E–PG neurons, there is an orderly and stereotyped mapping from the location of the dendrite of the cell to the location of its axon terminal in the protocerebral bridge<sup>47</sup>. Following previous studies<sup>3,4</sup>, we chose to image E–PG axons rather than dendrites because the axons are more superficial, and so more optically accessible. The imaging region was 256 × 128 pixels, and 8–12 slices deep in the z axis (3–5 µm per slice), resulting in a 6–9 Hz volumetric scanning rate.

Volumetric z-scanning signals from the piezoelectric objective scanner were acquired simultaneously with analogue output signals from the visual panorama and analogue outputs from FicTrac via a NiDAQ PCI-6341 at 4 kHz. Two-photon calcium imaging data were acquired using ScanImage 2018 (Vidrio Technologies) with National Instruments hardware provided by Vidrio (NI PXIe-6341).

### Measurement of locomotion

In all experiments except those shown in Fig. 3a, b, the fly stood on a 9-mm ball made of white foam (FR-4615, General Plastics) painted with black shapes. The ball floated above a plenum made of opaque ABS-like plastic (Figs. 1, 2, 3c, 5) or optically clear acrylic (Fig. 4) and was 3D-printed by Autotiv. Air was flowed into the plenum at the base and flowed out at the top in the semi-spherical depression that cradled the ball. The ball was illuminated by either an infrared LED (780 nm M780L3, Thorlabs) with a ground glass diffuser (DG10-220-MD, Thorlabs) (Figs. 1, 2, 3c, 5) or a round board 36 infrared LED lamp (SODIAL) (Fig. 4). The movement of the ball was tracked at approximately 60–70 Hz using a video camera (Firefly MV FMVU-03MTM, Point Grey) fitted with a Computar Macro zoom 0.3–1×, 1:4.5 lens (Figs. 1, 2, 3c, 5) or a Tamron

23FM08L 8-mm 1:1.4 lens (Fig. 4). In experiments in which we used a 360° visual panorama (Fig. 4), the image of the ball was reflected to the camera using a mirror (Thorlabs broadband dielectric mirror, 750–1,100 nm, BB1-E03) positioned below the ball. Machine vision software (FicTrac) converted the image of the ball to an estimate of the position of the ball in all three axes of rotation<sup>48</sup>. FicTrac was modified to send real-time analogue measurements of all three motion axes of the ball to a USB DAQ (USB-3101, Measurement Computing). For closed-loop experiments, the yaw-position voltage signal was used to update the azimuthal position of the visual cues displayed on the panorama.

### Visual panorama

Visual stimuli were presented using a circular panorama (IORodeo) composed of modular square panels<sup>49</sup>. Each square panel was an 8 × 8 array of LEDs (8 × 8 'pixels') that refreshed at 372 Hz or faster<sup>49</sup>. In electrophysiology experiments, these LEDs were green (peak = 525 nm). In imaging experiments, these LEDs were blue (peak = 470 nm) to minimize overlap with GCaMP6f emission. The vertical edge of the panorama was positioned approximately aligned with the vertical location of the fly. A single pixel along the top of the arena subtended around 3.6–3.7° of the visual field of the fly; this range of 0.1° is due to the fact that individual pixels within each flat 8 × 8 array have slightly different distances from the fly's eye. A single pixel at the bottom of the arena subtended around 2.7°. These differences in pixel size were not compensated for in our experiments.

In Figs. 1, 2, 3c, 5, we used a panorama composed of 9 × 2 panels. It spanned 270° azimuth and was oriented slightly asymmetrically so that it covered the azimuthal range from 127° left of the midline to 143° right of the midline. In Fig. 3a, we used a panorama composed of 6 × 2 panels that spanned 180° azimuth. In Fig. 4, we used a panorama composed of 12 × 2 panels that spanned 360° azimuth. All visual panoramas were the same height and spanned approximately 43° vertically within the visual field of the fly.

In electrophysiological experiments, to reduce electrical noise, the panorama was wrapped with a grounded copper mesh that was coloured with a black marker to reduce reflections. To further reduce reflections, the front surface of each panel was covered with a diffuser (SXF-0600 Snow White Light Diffuser, Decorative Films). In imaging experiments, instead of diffuser film, we used tracing paper as a diffuser, and four layers of filters (Rosco, R381, bandpass centre 440, full-width at half maximum of 40 nm) were used to minimize detection of the visual stimulus by the GCaMP6f emission collection channel.

### Open- and closed-loop modes of visual stimuli

To map visual receptive fields, we used a bright vertical bar (2 pixels wide, 7°) that spanned the full height of the panorama (around 43°). The bar was flashed for 500 ms followed by 500 ms of darkness. During open-loop mode, the display updated at 50 Hz. The bar was presented in a pseudorandom order at 35 different evenly spaced azimuthal positions across the screen (–120° to 135°). During each open-loop epoch, each bar position was used 4–5 times in total. For R neuron recordings, fewer positions were used (27 positions, –139° to 56°) and each location was used 5–6 times in total.

To map heading tuning curves and to provide visuomotor training (closed-loop mode), we used a visual panorama containing either one vertical bar (one-cue) or two bars positioned on opposite sides of the virtual world (two-cue). Each vertical bar was identical to the bar we presented in open-loop mode. In closed-loop mode, we controlled the azimuthal position of the visual pattern using the yaw-position voltage output from FicTrac. Between consecutive closed-loop epochs were 3–40 s of darkness, after which we shifted the pattern randomly (Fig. 4) or by a variable 45° or 90° increment (Figs. 2, 5) before returning to closed loop. Analogue output signals from the visual panel system and from FicTrac were digitalized with a NiDAQ PCI-6251 (National Instruments) at 20 kHz (electrophysiology) or with a NiDAQ PCI-6341

(National Instruments) at 4 kHz (calcium imaging). In Fig. 4, the 360° yaw output signal was mapped directly to the 360° visual panorama. In Figs. 2, 5, we needed to use a 270° panorama owing to the space constraints imposed by the electrophysiology set-up; therefore, the 360° yaw output signal was mapped linearly to the 270° panorama so that objects did not disappear when they reached one edge of the panorama but instead moved immediately across the gap<sup>1</sup>. Therefore, for example, whenever the fly made a 20° fictive right turn, the visual pattern would move 15° left. The exception to this is whenever the bar passed through the 90° gap; in that case, the bar traversed the gap immediately, as if the gap did not exist. How often this jump occurred varied from fly to fly depending on walking speed. We estimate that our most active flies experienced these 90° jumps of the cue around 10 times per minute during a typical one-cue closed-loop trial. Note that in the 270° panorama, the two-cue pattern contained two bars spaced 135° apart.

In pilot electrophysiology recordings, during closed-loop epochs, the 360° yaw output signal was mapped to 360° of visual space (rather than 270°). This meant that the visual cue was only displayed when it resided on the 270° panorama, and the cue simply disappeared when it moved into the 90° sector in which the panels were missing. The heading tuning data from these 16 recordings were not included in the final dataset, but some open-loop visual responses from these neurons are included in Fig. 1g. We did not observe any systematic differences in the open-loop visual responses of these neurons from pilot recordings.

## Optogenetic stimulation

Chrimson<sup>50</sup>-expressing flies were raised on cornmeal-agar medium supplemented with rehydrated potato flakes (Carolina Biological Supply) mixed with 100 µl of all-*trans*-retinal stock solution (Sigma; 17 mM in ethanol). Fly vials were wrapped in foil to prevent photo-conversion of the all-*trans*-retinal. Controls in Fig. 3b were raised on molasses food without all-*trans*-retinal. For optogenetic stimulation, we used the Hg-lamp source (U-LH100HG) to deliver a 5-ms pulse of green light (530–550 nm, 2–4 mW, TRITC–Cy3 filter cube, Chroma) through the objective. A shutter (Uniblitz Electronic) controlled the pulse duration.

## Experimental epoch structure

Each open-loop epoch always lasted 150 s and consisted of a sequence of random cue flashes. Each closed-loop epoch lasted 4 min (Figs. 2, 5) or 2 min (Fig. 4), during which time the visual pattern was continuously present and rotated in proportion to the fly's fictive yaw velocity. In Fig. 1, open-loop epochs were usually interleaved with 4-min one-cue closed-loop epochs, although occasionally two open-loop epochs were delivered consecutively. In Fig. 2, at least one 4-min one-cue closed-loop epoch was presented before obtaining a recording, and after the recording was obtained, open-loop epochs and 4-min one-cue closed-loop epochs were interleaved. In Fig. 3c, only open-loop epochs were presented. In Fig. 4, for pre-training, we presented at least five 2-min one-cue closed-loop epochs. For training, we presented ten 2-min two-cue closed-loop epochs. For post-training, we presented at least two 2-min one-cue closed-loop epochs. In Figs. 5, 1–6 epochs of one-cue closed-loop experience were presented before obtaining an E–PG neuron recording. Once the recording was obtained, the epoch structure was as follows. First, for pre-training, we cycled through 4-min one-cue closed-loop epochs alternating with open-loop epochs, for a total of 2–6 cycles. For training, we presented three consecutive 4-min two-cue closed-loop epochs (experimental condition) or three consecutive 4-min one-cue closed-loop epochs (matched control condition). For post-training, we presented one open-loop epoch. This protocol was followed in all training experiments in Fig. 5, with two exceptions. In one case, pre-training consisted of an open-loop epoch, followed by a closed-loop epoch, followed by another open-loop epoch (that is, 1.5 cycles through the normal pre-training procedure). In the other case, during the closed-loop epochs before obtaining the recording, the fly

experienced a different visual pattern that consisted of sparse randomly distributed single pixels (a 'star field' pattern), and this fly also received two consecutive open-loop epochs (instead of one) during pre-training.

## Immunohistochemistry

**MultiColor FlpOut.** In Extended Data Fig. 5, MultiColor FlpOut (MCFO) was used to identify the morphological types of R neurons labelled by *R20A02-Gal4*. MCFO immunostaining was performed essentially as described previously<sup>43</sup>. Primary incubation solution contained mouse anti-Bruchpilot (1:30, Developmental Studies Hybridoma Bank, nc82), rat anti-Flag (1:200, Novus Biologicals), rabbit anti-haemagglutinin (HA; 1:300, Cell Signaling Technologies) antibodies and 5% normal goat serum (NGS) in PBST. Secondary incubation solution contained Alexa Fluor 488-conjugated goat anti-rabbit (1:250, Invitrogen), ATTO 647-conjugated goat anti-rat (1:400, Rockland) and Alexa Fluor 405-conjugated goat anti-mouse (1:500, Invitrogen) antibodies and 5% NGS in PBST. Tertiary incubation solution contained DyLight 550-conjugated mouse anti-V5 (1:500, Bio-Rad) antibody and 5% normal mouse serum in PBST.

**Visualization of biocytin-filled neurons.** Brains containing biocytin-filled neurons were processed after electrophysiological recording using standard procedures. Primary incubation solution contained mouse anti-Bruchpilot antibody (1:30, Developmental Studies Hybridoma Bank, nc82), chicken anti-GFP antibody (1:1,000, Abcam), Alexa Fluor 568-conjugated streptavidin (1:1,000, Invitrogen) and 5% NGS in PBST. Secondary incubation solution contained Alexa Fluor 488-conjugated goat anti-chicken antibody (1:250, Invitrogen), Alexa Fluor 633-conjugated goat anti-mouse antibody (1:250, Invitrogen), Alexa Fluor 568-conjugated streptavidin (1:1,000, Invitrogen) and 5% NGS in PBST.

**Confocal microscopy and image analysis.** Brains processed for MCFO were imaged using an Olympus FV1000 confocal microscope. Series of between 50 and 100 optical sections (1.0-µm spacing) were imaged using either a UPLFLN 40×/1.3 NA oil-immersion lens or a PLAPON 60×/1.42 NA oil-immersion lens. R neuron MCFO clones were classified into 11 subtypes according to previously published methods<sup>6</sup> based on the consensus of two experts. Maximum intensity z-projections were rendered and adjusted using cropping and thresholding tools in Fiji (ImageJ) and assembled into figures using Illustrator (Adobe).

Confocal microscopy of brains processed for biocytin fills, or to assess expression of Kir2.1::eGFP within R neurons (Fig. 3), was performed using a Leica SP8 or Leica SPE equipped with a 40×/1.3 NA oil-immersion lens. Cell body counting of eGFP-labelled R neurons was performed independently by two experts using the Fiji Cell Counter plugin<sup>51</sup>, and the mean count for each brain hemisphere is reported (Extended Data Fig. 6).

## Data analysis

**Visual receptive fields of E–PG neurons.** In Figs. 1g, 2c, 3c, and 5b (and Extended Data Figs. 1–4, 8, 9), visual responses were calculated by taking the mean voltage during the final 250 ms of the 500-ms cue flash, and subtracting the mean voltage during the 250 ms preceding the flash, averaged over all presentations of the cue at each position. For display, visual receptive field curves were often smoothed using a median filter with a width of three cue positions (Figs. 1g, 3c and Extended Data Fig. 2) or two cue positions (Figs. 2c, 5b and Extended Data Figs. 3, 4, 8, 9). Peak visually evoked hyperpolarization (Fig. 3c) and mean visually evoked hyperpolarization (Extended Data Fig. 6) were calculated on the median-filtered tuning curves.

**Heading tuning of E–PG neurons.** In Figs. 2, 5 (and Extended Data Figs. 3, 4, 8, 9), heading tuning curves were calculated by first binning heading into 35 bins centred on the visual cue positions. The voltage



trace was filtered using a median filter with a width of 40 ms to remove spikes, and the mean-filtered voltage was measured for each heading during an epoch. For heading tuning curves calculated from multiple epochs, the voltage measurement for each heading bin was weighted relative to the number of samples in each individual epoch and the mean was then taken across epochs. For display, heading tuning curves were often smoothed using a median filter with a width of two cue positions (Figs. 2c, 5b and Extended Data Figs. 3, 4, 8, 9).

**Yaw during open-loop epochs.** In Fig. 1e, the FicTrac yaw-position signal was unwrapped, converted into radians, low-pass filtered (Butterworth) at 25 Hz and differentiated to obtain angular velocity. On rare occasions, a value of more than  $2500^\circ \text{ s}^{-1}$  occurred in an isolated time sample; this was probably due to imperfect nature of the unwrapping-and-differentiation procedure. These values were replaced with the value of the preceding sample. In Fig. 1f, the time-averaged yaw velocity was calculated by taking the mean yaw position during the final 250 ms of the flash and subtracting the mean yaw position during the 250 ms directly preceding the flash, and then dividing by the elapsed time (500 ms). We averaged data from left and right versions of the same cue displacement (because it seemed unlikely that a large group of flies would show a systematic bias in the right or left direction) to obtain mean yaw velocity responses to a total of 16 cue positions for each of 73 flies, thus obtaining  $73 \times 16$  data points. We took the mean across flies at each cue position and plotted this as the black line in Fig. 1f. Next, to model the null case (in which visual cue position has no effect), we randomly drew 73 values (with replacement) from the matrix, without regard for cue position or fly identity, and we calculated the mean of these 73 values; we constructed a bootstrap distribution by repeating this procedure 10,000,000 times, each time calculating the mean of 73 randomly drawn values. This bootstrap distribution was used to obtain a 95% confidence interval, which was then adjusted for multiple comparisons using a Bonferroni correction ( $m = 16$  tests). None of the true mean values (black) were outside this adjusted confidence interval (magenta lines). We used the same procedure in Extended Data Fig. 1e, except that the independent variable was the distance of the cue jump rather than the position of the cue. Finally, as a further control, we also examined whether any individual flies had a significant yaw velocity response to any cue position (Extended Data Fig. 1d). Because individual flies might be right- or left-handed<sup>52</sup>, we did not average data from right and left cue positions in this analysis; thus there were 35 cue positions. For each fly, we computed trial-averaged yaw velocity for each of 2–8 open-loop epochs, and we created a matrix containing all cue positions for every epoch in the dataset of that fly. We then randomly drew a number of values (with replacement) from the matrix (number of epochs  $\times$  35 cue positions) to match the number of epochs that we recorded for that fly. This procedure was randomized with respect to cue position and epoch number. For each fly, a bootstrap distribution was obtained by repeating this procedure 100,000 times, each time calculating the mean of the drawn values. The difference between the observed trial-averaged yaw responses for each cue position and the mean of the bootstrap distribution was used to obtain a  $P$  value (two-sided). In this manner, a  $P$  value was calculated for every fly at every cue position ( $73 \times 35 P$  values). The statistical significance of each trial-averaged yaw was assessed for each fly and each position at with  $\alpha = 0.05$  using the Bonferroni–Holm method to correct for multiple comparisons. No tests showed a statistically significant yaw velocity for any individual fly at any cue position.

**Correlations between visual receptive fields and heading turning curves in E–PG neurons.** In Fig. 2c, d, heading tuning curves and visual receptive fields were smoothed using a median filter with a width of two cue positions. Correlation coefficients were computed on smoothed curves (40 pairs in total). In Fig. 2d, as a control, we randomly drew (with replacement) 40 heading tuning curves and 40 visual response

curves, yielding 40 correlation coefficients. The mean of that correlation value was then recorded. This process was repeated 10,000,000 times to build a bootstrap distribution and the 95% confidence interval of this distribution was computed.

**Visual receptive fields of R neurons.** In Fig. 3a, spikes were detected after low-pass filtering the recorded current at 1 kHz by identifying deflections greater than 15 pA that occurred outside a 0.5-ms refractory period. The spike rate was measured over the 500-ms visual stimulus period.

**Responses of E–PG neurons to optogenetic stimulation of R neurons.** In Fig. 3b, peak hyperpolarization was calculated as the trial-averaged voltage during a 1-s baseline period minus the minimum trial-averaged voltage reached in the 1 s following the 5-ms optogenetic stimulus. Four optogenetic stimulus trials were recorded per cell.

**E–PG ensemble representations of heading direction.** In Fig. 4, rigid motion correction in the  $x$ ,  $y$  and  $z$  axes was performed for the volumetric imaging stacks for every epoch using the NoRMCorre algorithm<sup>53</sup>. This algorithm performs piece-wise rigid registration of small overlapping sectors within the field of view, and then merges the sectors via interpolation, allowing approximate cancellation of non-rigid brain movement artefacts. Motion correction was parallelized on a high-performance computing cluster. For each epoch, we defined 16 regions of interest, corresponding to the 16 glomeruli in the protocerebral bridge; each region of interest was defined in one  $z$  plane. To calculate the time-dependent change in fluorescence ( $\Delta F/F$ ) for each glomerulus, we used a baseline fluorescence ( $F$ ) defined as the mean of the lowest 5% of raw fluorescence values across the entire experiment for that glomerulus. We excluded from the baseline the rare frames that were lost as a result of the rigid motion correction algorithm. The singular bump of activity in E–PG dendrites within the ellipsoid body<sup>1</sup> translates into two bumps in the protocerebral bridge<sup>3,4</sup>; these two bumps move together, so that the signal has a spatial period of eight glomeruli in the protocerebral bridge. Therefore, to calculate the neural representation of heading direction, we took the spatial Fourier transform of  $\Delta F/F$  in the protocerebral bridge across all 16 glomeruli. We used the phase of the Fourier component at eight glomeruli as the phase of the neural representation of heading for each time point; this procedure was described previously<sup>3</sup>. We used the sign convention in which a positive change in phase corresponds to a rightward movement of the bumps in the protocerebral bridge, and a clockwise movement of the bump in the ellipsoid body (when viewed from the posterior side of the brain). For display purposes only, in Fig. 4a, we averaged the  $\Delta F/F$  signals from the right and left half of the protocerebral bridge (which is why only one bump is visible); this averaging was not performed as part of the data analyses described above.

**Offset of the E–PG ensemble reference frame.** In Fig. 4c, d, to calculate the offset of the reference frame (the difference between the fly's heading and the neural representation of heading), we first downsampled the behavioural data to match the volumetric imaging rate (6–9 Hz). We removed time points in which the FicTrac analogue signals were problematic or when the power of the Fourier transform was below a specified threshold (0.1). We also excluded the first 3 s of each 2-min closed-loop epoch due to a delay between imaging trigger and the start of the visual stimulus. We then took the angular position of the visual panorama from the analogue voltage output of the LED panel system (positive defined as to the right of the fly, or clockwise when viewed from above the set-up). We calculated the offset of the E–PG ensemble reference frame as the negative of the spatial Fourier transform phase minus the position of the visual panorama. This value is consistent with previously published methods<sup>3</sup> to calculate the offset between the bump position in the protocerebral bridge and the ball yaw

position. To quantify offset probability (Fig. 4d, f and Extended Data Fig 7), we analysed the final 4 min of each pre-training block, the full 20 min of each training block, and the first 4 min of each post-training block.

**Effect of training on visual receptive fields.** In Fig. 5 (and Extended Data Figs. 8, 9), heading tuning curves and visual receptive fields were smoothed using a median filter with a width of two samples (cue positions). The last pre-training open-loop epoch (probe 1) and the first post-training open-loop epoch (probe 2) were used for the following analyses. In Fig. 5d–f, the absolute change was obtained by subtracting the two visual receptive fields (post-training minus pre-training), then summing the absolute value of the difference ( $\Delta$ ) over all cue positions, and finally dividing by the number of positions. In Fig. 5d, the change in receptive field shape was obtained by cross-correlating probe 1 and probe 2 and calculating  $1 - R^2$ . In Fig. 5f, the modulation by heading during turning was taken as the difference between the maximum and the minimum of the heading direction turning curve.

**Controls for training.** In Fig. 5e, to estimate the drift in visual receptive field under control conditions, we had flies navigate in a one-cue world (rather than a two-cue world) during the waiting period between the open-loop epochs. In some cases (matched control in Extended Data Fig. 9), flies received exactly the same protocol as the experimental condition except with one-cue closed-loop during the training period; in other words, these matched controls received 12 consecutive minutes of one-cue (rather than two-cue) closed-loop epochs during the ‘training’ period. For other controls (control in Extended Data Fig. 9), we identified experiments from Fig. 2 in which the recording had lasted long enough for us to present four open-loop epochs interleaved with four one-cue closed-loop epochs. In these recordings, the second and fourth epochs were separated by more than 12 min (typically around 15 min) and so they are appropriate controls for the training protocol. We therefore treated the second and fourth open-loop epochs as if they were ‘probe 1’ and ‘probe 2’ epochs in a training experiment, and we analysed them as described above for the true training experiments. The important distinction is that this second group of control flies experienced one-cue rather than two-cue closed-loop epochs during the window between probe 1 and probe 2.

## Data inclusion

We include epochs for Figs. 1, 2 if the cell was healthy; specifically, this meant that the epoch-averaged voltage was below  $-33$  mV and within 15 mV of the voltage observed at the start of the first epoch of the experiment, and if the spike amplitude was more than 50% of the amplitude observed in the first epoch. Closed-loop epochs were included if the fly visited all heading directions during that epoch. Cells were included if  $\geq 2$  open-loop epochs met these criteria; in Fig. 2 we also required that  $\geq 2$  closed-loop epochs met these criteria. In Fig. 3, cells were included if  $\geq 2$  open-loop epochs met our cell health criteria. A single recording from the *UAS/+* control genotype was excluded because the biocytin fill showed that it was not an E–PG neuron. All other biocytin-filled neurons analysed during this project (that is, 65 out of 66 neurons) were confirmed to be E–PG neurons. All recordings that were not imaged post hoc were therefore assumed to target E–PG cells. We excluded 5 out of 24 flies in Fig. 4 owing to either weak fluorescence or an unstable offset between the angle of the E–PG bump and the fly’s heading angle at the end of the initial closed-loop one-cue epoch. In Fig. 5, cells were included if the epoch-averaged voltage from all epochs of the experiment (pre-training, training, post-training) was below  $-33$  mV and if the fly visited all heading directions during the two epochs (8 min) of one-cue closed-loop before training and during the final two epochs (8 min) of two-cue closed-loop training. We required that the fly’s mean yaw velocity was  $>20^\circ/\text{s}$  during the final 2 epochs of the two-cue closed-loop training; 10 cells were excluded due to this restriction.

We also removed recordings in which the visual receptive field and/or heading turning curve were almost flat during the pre-training period ( $\text{max} - \text{min} \leq 2$  mV); six cells were removed due to this restriction.

On occasion, during E–PG neuron electrophysiological recordings, we observed unexpected large inhibitory postsynaptic potentials with a stereotyped sharp onset, a large amplitude ( $>15$  mV) and a stereotyped time course. They were followed by a prolonged period of depolarization when the variance of the voltage trace was also diminished. These events interfered with visual and heading tuning measurements; therefore, for Figs. 1–3, any epoch in which such an event occurred was excluded from the analysis. For Fig. 5, the event was clipped but the rest of the epoch was used; 5% of open loop epochs and 10% of closed-loop epochs were clipped in this manner.

Analysis was performed using MATLAB R2016b, R2017a and R2017b (MathWorks).

## Determination of sample sizes

For genetic perturbation experiments (Fig. 3c), the number of experiments performed was determined by first collecting a pilot dataset of ( $n = 4$  for the three genotypes using the *R20A02-Gal4* driver line). On the basis of the initial effect size, power analysis was used to determine the number of experiments needed to test the hypothesis that visually evoked hyperpolarization was smaller in the experimental genotype. For all other experiments, sample sizes were chosen based on standard sample sizes in the field.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

## Code availability

Analysis code is available at [https://github.com/wilson-lab/FisherLu-DAlessandroWilson\\_AnalysisCode](https://github.com/wilson-lab/FisherLu-DAlessandroWilson_AnalysisCode).

37. Jenett, A. et al. A GAL4-driver line resource for *Drosophila* neurobiology. *Cell Rep.* **2**, 991–1001 (2012).
38. Wang, J., Zugates, C. T., Liang, I. H., Lee, C. H. & Lee, T. *Drosophila* Dscam is required for divergent segregation of sister branches and suppresses ectopic bifurcation of axons. *Neuron* **33**, 559–571 (2002).
39. Pfeiffer, B. D. et al. Refinement of tools for targeted gene expression in *Drosophila*. *Genetics* **186**, 735–755 (2010).
40. Hoopfer, E. D., Jung, Y., Inagaki, H. K., Rubin, G. M. & Anderson, D. J. P1 interneurons promote a persistent internal state that enhances inter-male aggression in *Drosophila*. *eLife* **4**, e11346 (2015).
41. Hardie, R. C. et al. Calcium influx via TRP channels is required to maintain PIP2 levels in *Drosophila* photoreceptors. *Neuron* **30**, 149–159 (2001).
42. Chen, T. W. et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
43. Nern, A., Pfeiffer, B. D. & Rubin, G. M. Optimized tools for multicolor stochastic labeling reveal diverse stereotyped cell arrangements in the fly visual system. *Proc. Natl Acad. Sci. USA* **112**, E2967–E2976 (2015).
44. Goodman, M. B. & Lockery, S. R. Pressure polishing: a method for re-shaping patch pipettes during fire polishing. *J. Neurosci. Methods* **100**, 13–15 (2000).
45. Green, J., Vijayan, V., Mussells Pires, P., Adachi, A. & Maimon, G. Walking *Drosophila* aim to maintain a neural heading estimate at an internal goal angle. Preprint at <https://doi.org/10.1101/315796> (2018). If ref. 45 (preprint) has now been published in final peer-reviewed form, please update the reference details if appropriate. If ref. 45 has now been accepted or published in peer-reviewed form, please update the reference.
46. Gouwens, N. W. & Wilson, R. I. Signal propagation in *Drosophila* central neurons. *J. Neurosci.* **29**, 6239–6249 (2009).
47. Wolff, T., Iyer, N. A. & Rubin, G. M. Neuroarchitecture and neuroanatomy of the *Drosophila* central complex: a GAL4-based dissection of protocerebral bridge neurons and circuits. *J. Comp. Neurol.* **523**, 997–1037 (2015).
48. Moore, R. J. et al. FicTrac: a visual method for tracking spherical motion and generating fictive animal paths. *J. Neurosci. Methods* **225**, 106–119 (2014).
49. Reiser, M. B. & Dickinson, M. H. A modular display system for insect behavioral neuroscience. *J. Neurosci. Methods* **167**, 127–139 (2008).

50. Klapoetke, N. C. et al. Independent optical excitation of distinct neural populations. *Nat. Methods* **11**, 338–346 (2014).
51. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
52. Buchanan, S. M., Kain, J. S. & de Bivort, B. L. Neuronal control of locomotor handedness in *Drosophila*. *Proc. Natl Acad. Sci. USA* **112**, 6700–6705 (2015).
53. Pnevmatikakis, E. A. & Giovannucci, A. NoRMCorre: an online algorithm for piecewise rigid motion correction of calcium imaging data. *J. Neurosci. Methods* **291**, 83–94 (2017).

**Acknowledgements** We thank D. Anderson, T. Clandinin, B. Pfeiffer, G. Rubin and J. Tuthill for providing fly stocks; T. Clandinin, B. Bean, J. Drugowitsch, D. Ginty and members of the Wilson laboratory for providing feedback on the manuscript and J. Drugowitsch for providing advice on data analysis; G. Maimon for sharing designs for a fly holder and modified FicTrac software; O. Mazor and P. Gorelik at the Harvard Medical School Research Instrumentation Core (NEI Core Grant for Vision Research EY012196) for their help constructing the virtual-reality systems. This work was supported by the Harvard Neurobiology Imaging Facility (NINDS P30 NS072030). This work was funded by NIH awards U19NS104655, F30DC017698 (to J.L.) and

T32GM007753 (to J.L.). Y.E.F. is supported by a HHMI Hanna H. Gray Fellowship. R.I.W. is an HHMI Investigator.

**Author contributions** Y.E.F., J.L. and R.I.W. designed the study. Y.E.F. performed and analysed electrophysiology experiments. J.L. performed and analysed two-photon calcium-imaging experiments. I.D. performed and analysed confocal-microscopy imaging experiments. Y.E.F. and R.I.W. wrote the manuscript with input from J.L. and I.D.

**Competing interests** The authors declare no competing interests.

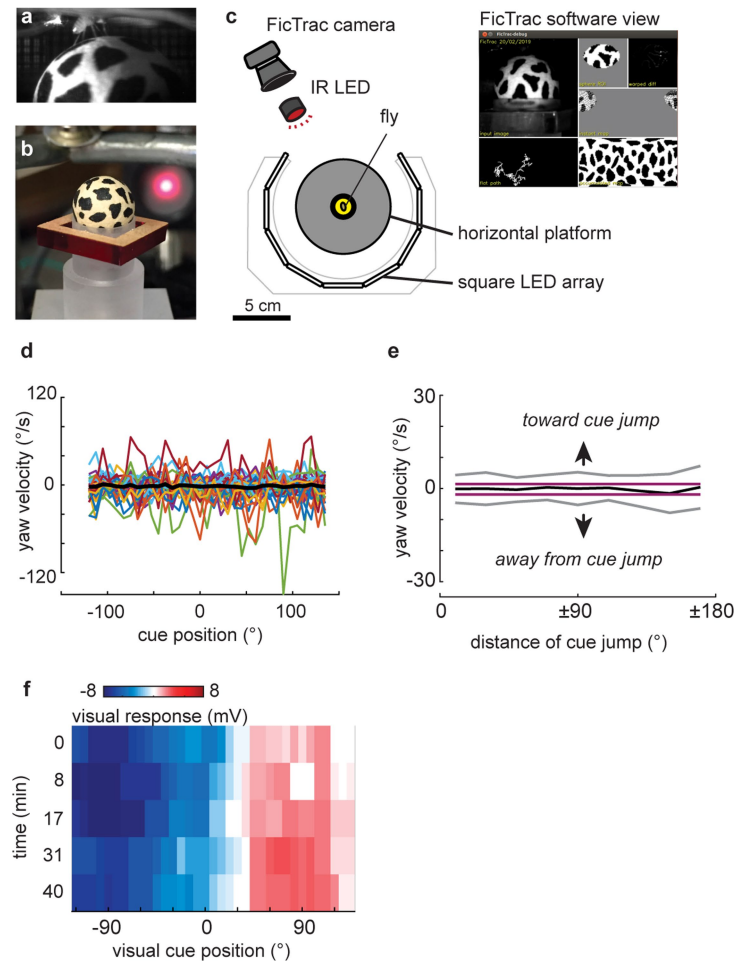
#### **Additional information**

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1772-4>.

**Correspondence and requests for materials** should be addressed to R.I.W.

**Peer review information** *Nature* thanks Lisa Giocomo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

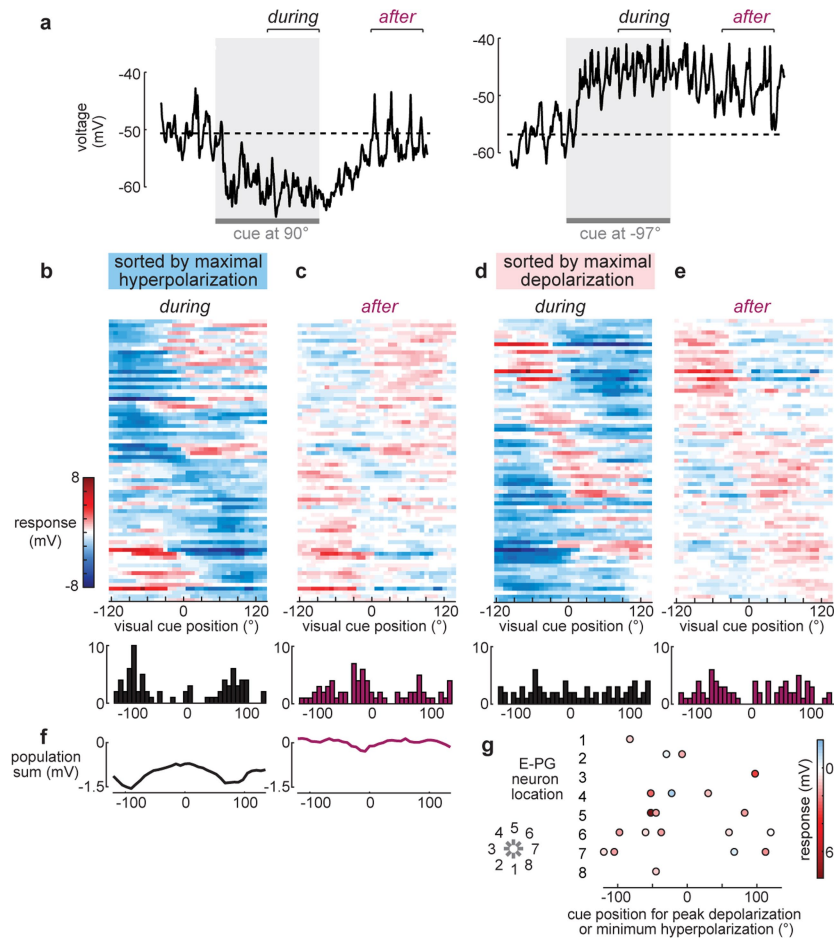
**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



#### Extended Data Fig. 1 | Measuring behaviour and E-PG visual responses.

**a**, Side view of a fly walking on an air-cushioned ball during an electrophysiology experiment. **b**, Image of the ball and plastic holder. Air flows up through the holder and out the semi-spherical depression that cradles the ball. **c**, Schematic of the experimental set-up viewed from above. The fly is secured in an aperture in the centre of a horizontal platform. The platform is surrounded by a circular panorama. The panorama is composed of square LED arrays<sup>49</sup> (2 squares vertically  $\times$  12 squares horizontally). The ball is illuminated by an infrared (IR) LED, which is visible as a red spot in **b**. A camera captures an image of the ball to enable tracking using FicTrac<sup>48</sup>. Inset shows FicTrac view. Camera and infrared LED are not drawn to scale. **d**, The yaw velocity of the fly compared to the cue position. This is the dataset that is the basis for Fig. 1f, but here broken down into averages for each individual fly, and with right (+) and left (–) cue positions kept separate. Positive velocities are right turns, and negative velocities are left turns. No tests showed a statistically significant yaw velocity ( $P < 0.05$ , two-sided comparison to bootstrap distribution) for any individual fly at any cue position. For details of analysis, see Methods, ‘Yaw

during open-loop epochs’. **e**, Yaw velocity in response to the visual cue presentation. This analysis is the same as that shown in Fig. 1f, but here yaw velocity is plotted against the distance of the cue jump between consecutive trials. As in Fig. 1f, we show mean (black)  $\pm$  1 s.d. (grey) across experiments (73 experiments in 68 flies). Magenta lines show the bootstrapped 95% confidence interval of the mean across flies after randomizing cue positions, Bonferroni-corrected for multiple comparisons. Because the mean lies within these bounds, it is not significantly different from random. This analysis further supports the conclusion that there is no systematic yaw response to the random flashes of the vertical bar. For details of analysis, see Methods, ‘Yaw during open-loop epochs’. **f**, The visual receptive field of an example cell measured multiple times over the course of a 40-min recording. Each row shows data from a separate visual mapping epoch. Data from this example cell are also shown in Fig. 1e. Note the stability of the visual receptive field over this time period. For experiments shown in this figure, we used *UAS-mCD8::GFP/UAS-mCD8::GFP;R60D05-Gal4/R60D05-Gal4* flies.

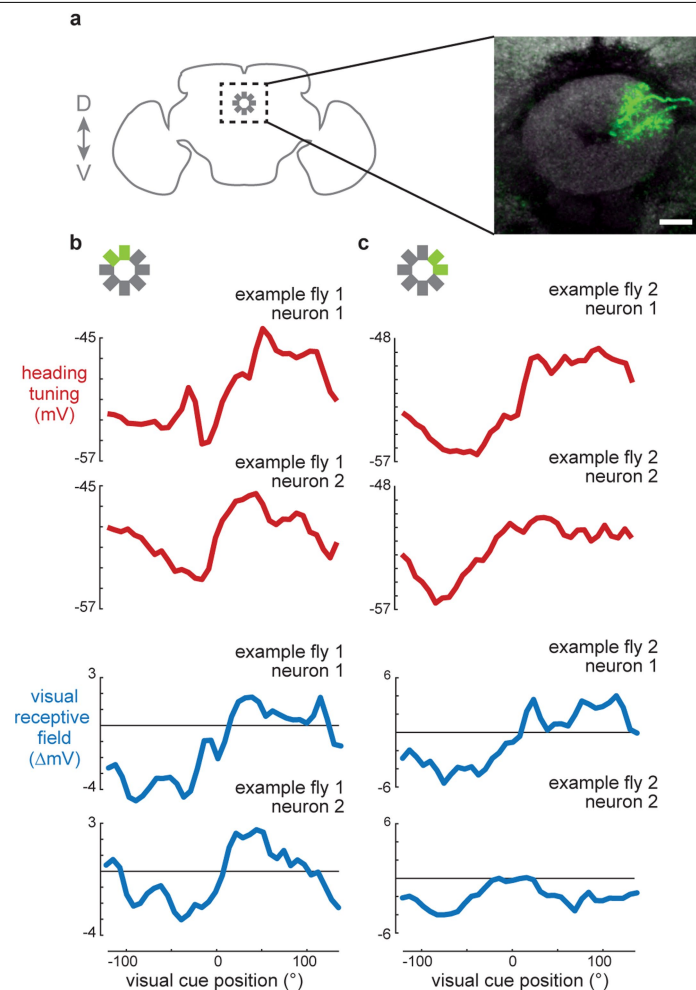


### Extended Data Fig. 2 | Visually evoked hyperpolarization and

**depolarization, during and after cue presentation.** **a**, Example voltage responses of the same E-PG neuron to two cue positions. Dashed lines indicate the mean baseline voltage before the cue. This neuron is hyperpolarized by the cue at  $90^\circ$  and depolarized by the cue at  $-97^\circ$ . Note that hyperpolarization decays more rapidly than depolarization. In **b**, to quantify visual receptive fields, we measured the change in voltage during cue presentation and after cue removal in the 250-ms windows marked in **a** with brackets, in both cases relative to baseline. **b**, Summary of E-PG visual receptive fields measured during cue presentation. Cells are sorted by the cue position that evokes maximal hyperpolarization. The histogram shows the number of E-PG neurons with maximal hyperpolarization at each cue position (73 E-PG neurons in 68 flies). **c**, Summary of E-PG visual receptive fields measured after cue removal. Cell order is the same as in **b**. Note that hyperpolarizing responses tend to decay, whereas depolarizing responses tend to persist; this is consistent with the hypothesis that the hyperpolarization during cue presentation is due to

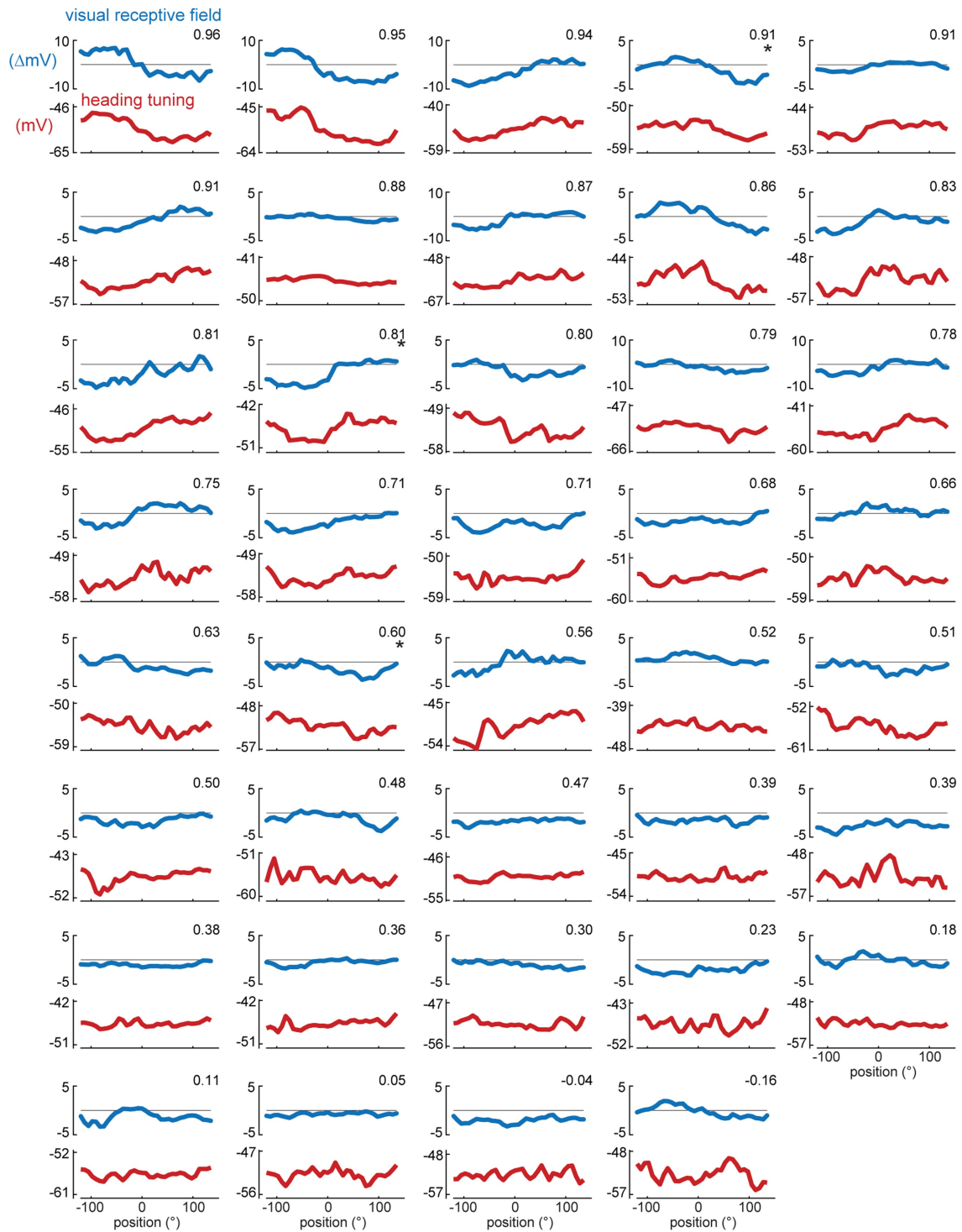
direct synaptic inhibition from R neurons, whereas depolarization is polysynaptic and caused by withdrawal of tonic synaptic inhibition. The histogram shows the number of E-PG neurons with maximal hyperpolarization after cue removal for each cue position. **d**, Same as **b**, but sorted by the cue position that evoked maximal depolarization (minimal hyperpolarization), as in Fig. 1g. **e**, Same as **c**, but with the cell order as in **d**. **f**, Summed response across all neurons measured during (left) and after (right) the cue. The left curve has a pair of minima around  $\pm 100^\circ$ ; this bias is probably inherited from R neuron receptive fields, which are biased towards positions offset from the visual midline<sup>5</sup>. By contrast, the right curve is relatively flat. **g**, Visual cue position eliciting maximal depolarization (minimum hyperpolarization), plotted versus E-PG neuron location, for the 21 recorded E-PG neurons that were filled. No significant correlation was observed (circular correlation coefficient =  $-0.15$ ,  $P = 0.49$ )<sup>36</sup>. For experiments shown in this figure, we used *UAS-mCD8::GFP/R60D05-Gal4/R60D05-Gal4* flies.





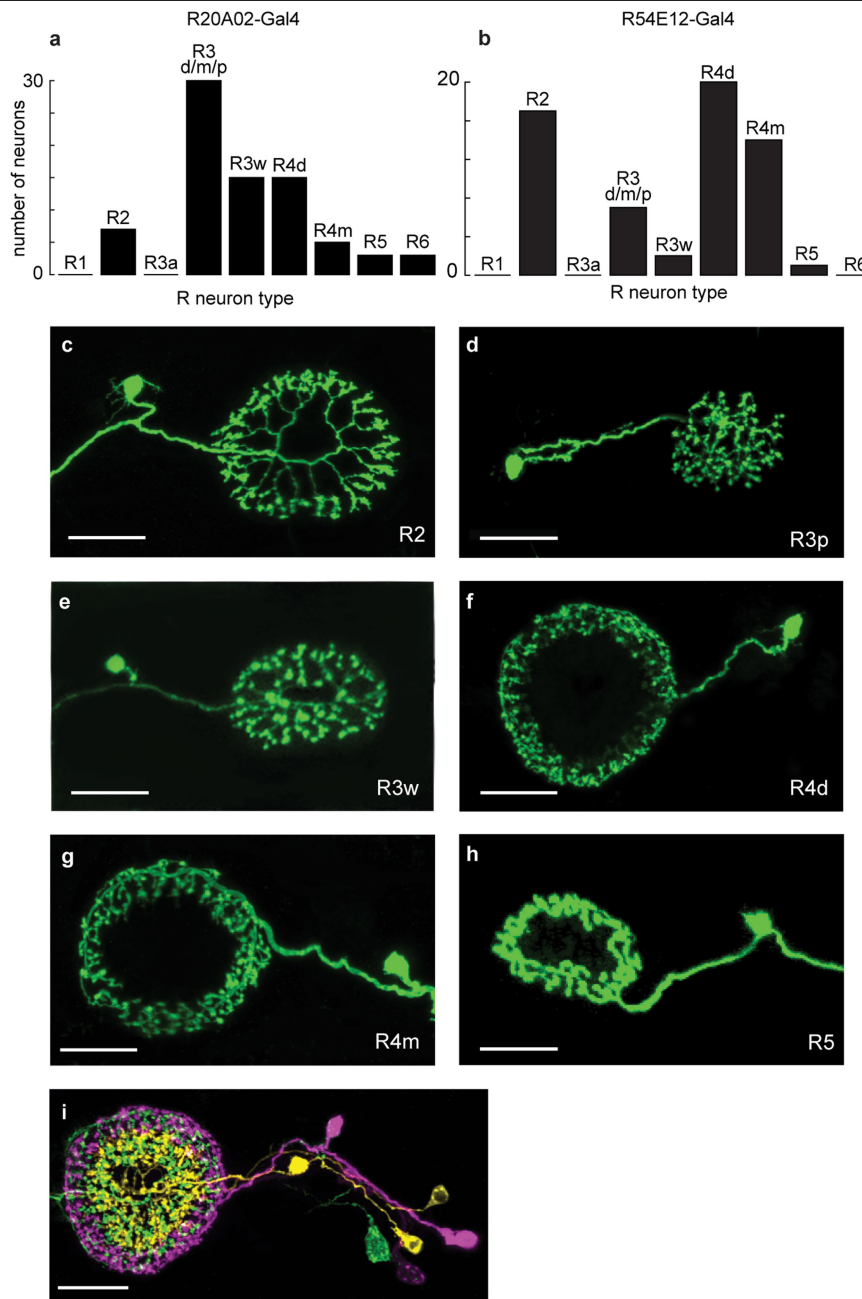
**Extended Data Fig. 3 | E-PG neuron pairs recorded sequentially from the same brain.** **a**, Two biocytin filled dendrites (green) from sequentially recorded E-PG neurons that innervate adjacent wedges within the ellipsoid body. Neuropil reference marker is shown in grey (anti-nc82 antibody). Images are maximum intensity z-projections. Scale bar, 10  $\mu$ m. The schematic shows the approximate position of ellipsoid body and E-PG dendrites from a coronal view of the fly brain. **b, c**, Heading tuning (red, measured in VR) and visual receptive field (blue, measured with random flashes) from sequentially recorded E-PG pairs from two example flies. Dendritic locations of the recorded neurons are green in the ellipsoid body schematic above each set of

plots. In both cases, by chance, the two dendrites were physically adjacent. In both cases, adjacent E-PG neurons from the same fly exhibited similar visual receptive fields and heading tuning curves, supporting the conclusion that adjacent E-PG cells typically receive inhibition from adjacent regions of visual space and represent adjacent heading directions. Comparing the visual receptive field and the heading tuning curve for each neuron yielded correlation coefficients (Pearson's) of 0.76 (fly 1 neuron 1), 0.90 (fly 1 neuron 2), 0.95 (fly 2 neuron 1) and 0.65 (fly 2 neuron 2). For experiments shown in this figure, we used *UAS-mCD8::GFP/UAS-mCD8::GFP; R60D05-Gal4/R60D05-Gal4* flies.



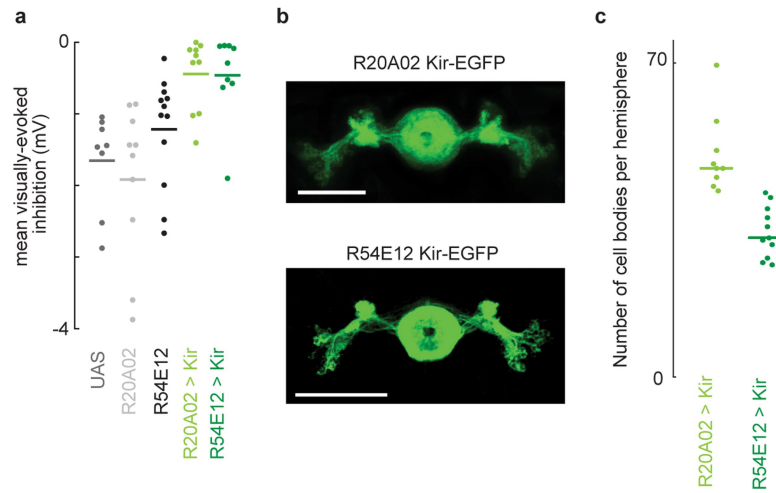
**Extended Data Fig. 4 | Visual receptive fields and heading tuning of E-PG neurons.** Heading tuning (red, closed-loop mode) and visual receptive fields (blue, open-loop mode) for all 40 recorded E-PG neurons (from 39 flies). For each neuron, the correlation coefficient (Pearson's) is reported for the

comparison between the visual receptive field and the heading tuning curve. Asterisks denote data also shown in Fig. 2. For experiments shown in this figure, we used *UAS-mCD8::GFP/UAS-mCD8::GFP;R60D05-Gal4/R60D05-Gal4* flies.



**Extended Data Fig. 5 | R neuron types labelled by *R20A02-Gal4* and *R54E12-Gal4* described by MCFO.** **a**, Observed numbers of R neurons belonging to each type from a dataset of  $n = 78$  single-neuron MCFO clones<sup>43</sup> from the *R20A02-Gal4* line. R neuron types were classified according previously published methods<sup>6</sup>. **b**, Same as in **a** but for the *R54E12-Gal4* line ( $n = 61$  single-neuron MCFO clones). **c–h**, Examples of single R neuron MCFO clones. Images are maximum intensity z-projections. Background labelling was manually

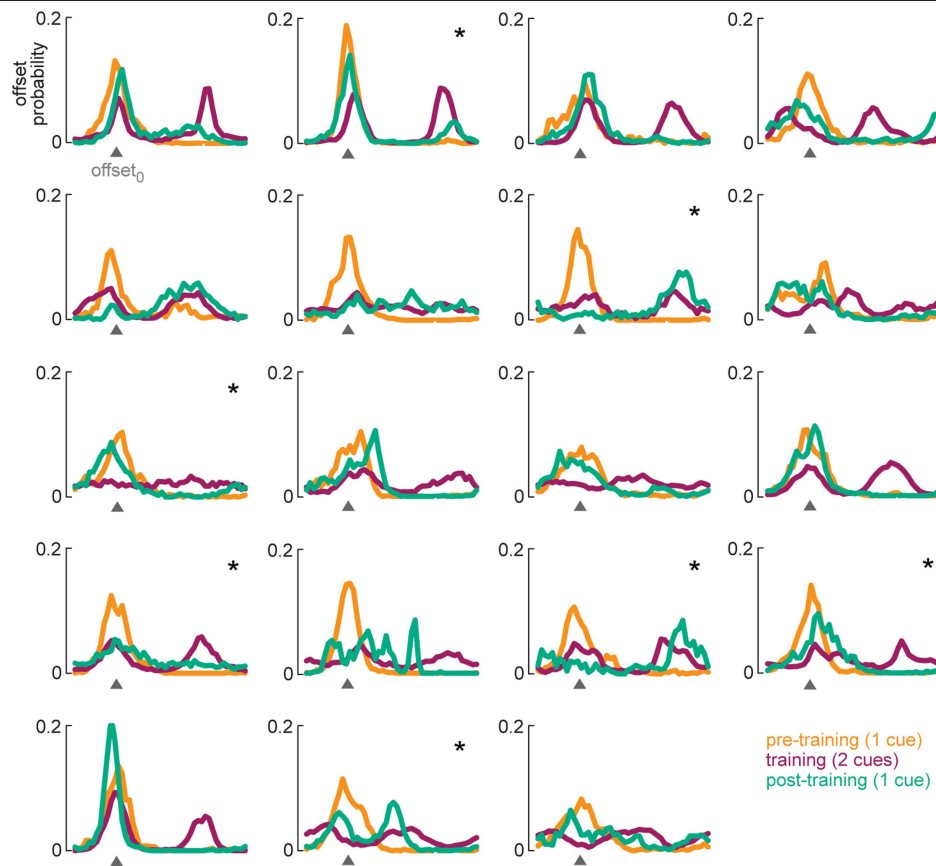
removed to improve clarity of specific neuronal morphologies. **i**, Multiple R neuron MCFO clones labelled in different colours using the *R20A02-Gal4* line. Image is a maximum-intensity z-projection. Scale bars, 20  $\mu\text{m}$ . For experiments shown in this figure, we used *R57C10-FLPGS.PEST; UAS(FRT.stop)myr::smGdP-HA, UAS(FRT.stop)myr::smGdP-V5, UAS(FRT.stop)myr::smGdP-Flag/R20A02-Gal4, R57C10-FLPGS.PEST; UAS(FRT.stop)myr::smGdP-HA, UAS(FRT.stop)myr::smGdP-V5, UAS(FRT.stop)myr::smGdP-Flag/R54E12-Gal4* flies.



**Extended Data Fig. 6 | Suppressing R neuron activity with two independent driver lines reduces visually evoked hyperpolarization in E-PG neurons.**

**a**, Same as Fig. 3c, except instead of measuring peak visually evoked hyperpolarization, we measured mean visually evoked hyperpolarization (by zeroing all non-negative visual responses and then averaging visual responses across all cue positions). From left to right:  $n = 8, 10, 12, 10, 9$ . Both Kir2.1 means are significantly different from corresponding genetic controls using two-sided Wilcoxon rank-sum tests. *R20A02 Kir2.1* versus *R20A02/+* and *UAS/+* ( $P = 0.0013$  and  $P = 0.0003$ , respectively), *R54E12 Kir2.1* versus *R54E12/+* and *UAS/+* ( $P = 0.005$  and  $P = 0.0025$ , respectively). **b**, R neuron population labelled by Kir2.1::eGFP. Images are maximum intensity z-projections. **c**, Numbers of R neurons per hemisphere expressing Kir2.1::eGFP in each experimental genotype,  $n = 9$  (*R20A02*) and  $n = 11$  (*R54E12*) (horizontal lines are means). On the basis of the previously reported total number of R neurons of each type<sup>6</sup> and our MCFO quantification of the R neuron types labelled by *R20A02-Gal4* and *R54E12-Gal4* (Extended Data Fig. 5), these cell counts suggest that *R20A02-*

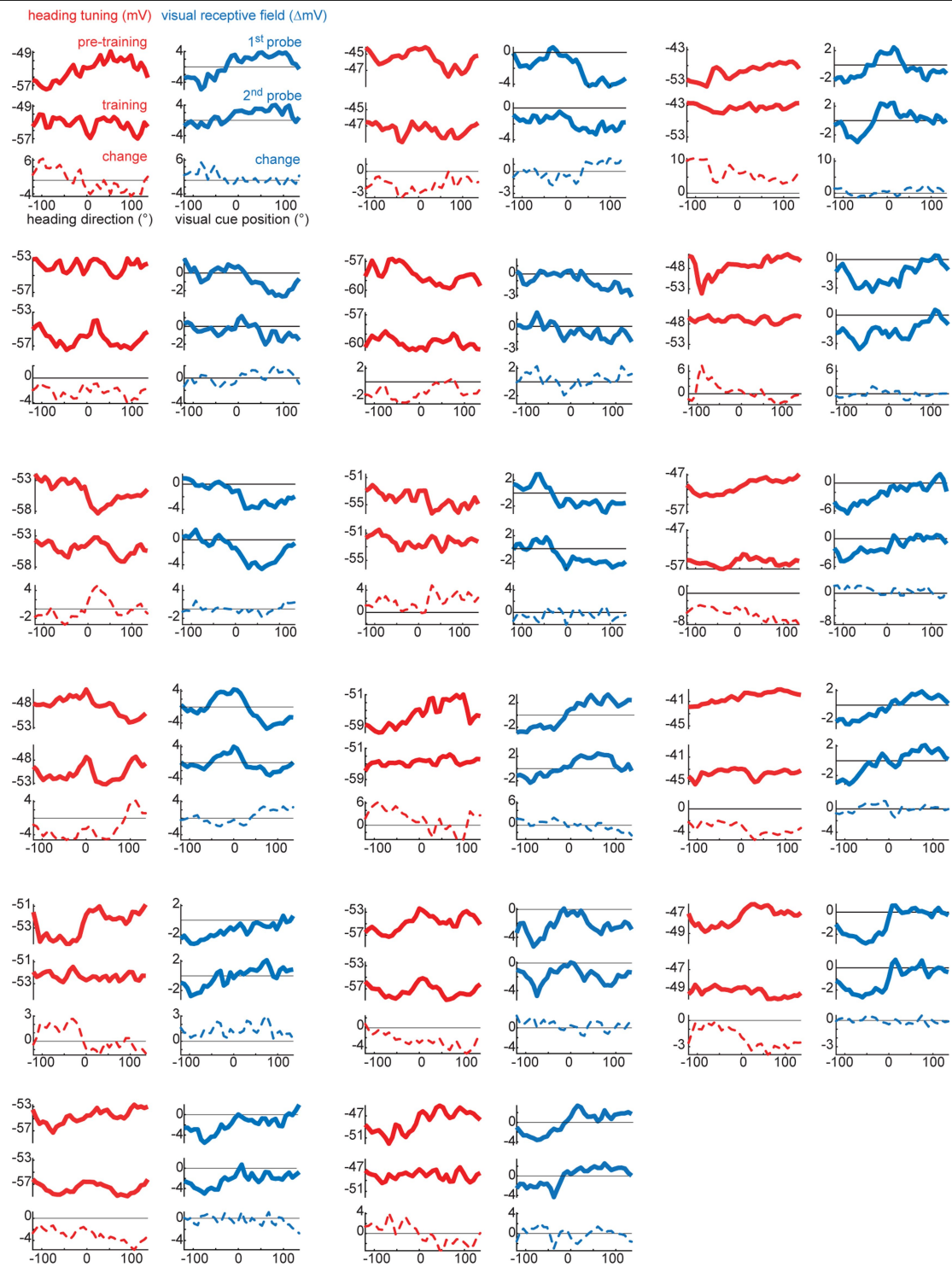
*Gal4* targets approximately 20% of R2, 30% of R4m and all R4d neurons. These counts suggest that *R54E12-Gal4* targets approximately 40% of R2 neurons and all R4m and R4d neurons. This incomplete targeting of outer R neurons may provide one explanation for the remaining visually evoked inhibition observed in some recordings (Fig. 3). Note that although both driver lines label other neurons in the central brain and visual system, R neurons appear to be the only cell type that is labelled by both lines. In the visual system, the driver line *R20A02-Gal4* targets one medulla intrinsic neuron, probably Mi12, and one cell type that arborizes in around layers 4–6 of the lobula, whereas the driver line *R54E12-Gal4* appears to target the medulla neuron Tm3. For experiments shown in this figure, we used *+w; R60D05-LexA/LexAop-mCD8::GFP; +/UAS-Kir2* (UAS-only control); *+w; R60D05-LexA/LexAop-mCD8::GFP; R20A02-Gal4/+* (*R20A02 Gal4*-only control); *+w; R60D05-LexA/LexAop-mCD8::GFP; R54E12-Gal4/+* (*R54E12 Gal4*-only control); *+w; R60D05-LexA/LexAop-mCD8::GFP; R20A02-Gal4/UAS-Kir2.1* (*R20A02 Kir2.1*); and *+w; R60D05-LexA/LexAop-mCD8::GFP; R54E12-Gal4/UAS-Kir2.1* (*R54E12 Kir2.1*) flies.



**Extended Data Fig. 7 | Offset probability histograms in training experiments.** Offset probability histograms during each segment of the training experiments shown in Fig. 4, for all 19 GCaMP imaging experiments (in 19 flies). As in Fig. 4, the circular mean during the pre-training period is defined

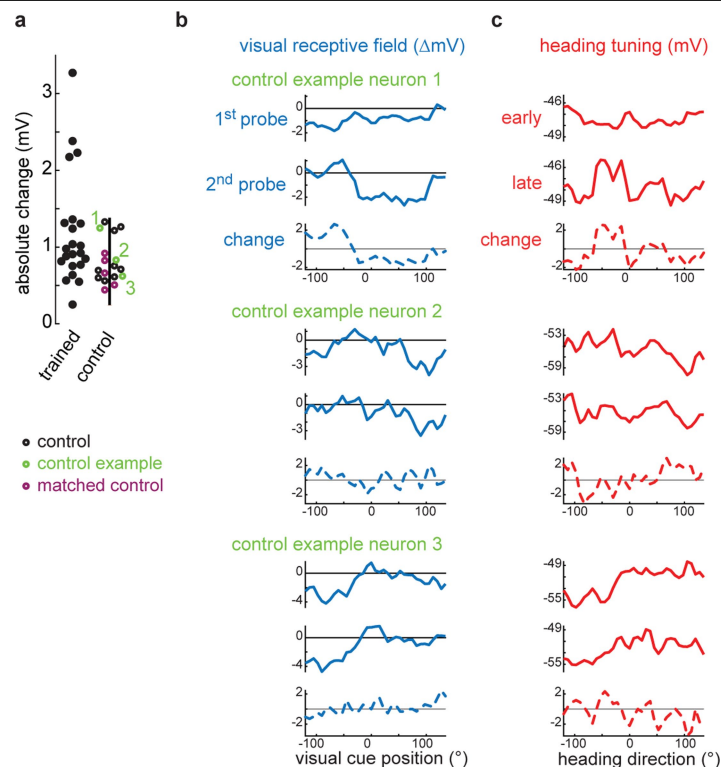
as  $\text{offset}_0$  (here marked with an arrowhead), and for display purposes we horizontally aligned all of the  $\text{offset}_0$  values in different flies. Asterisks mark data shown in Fig. 4. For experiments shown in this figure, we used  $+w; \text{UAS-GCaMP6f}/+; \text{R60D05-Gal4}/+$  flies.





**Extended Data Fig. 8 | Heading tuning and visual receptive field measurements in training experiments.** Heading tuning curves and visual receptive fields for all additional 17 E-PG neurons (from 17 flies) from the training experiments in Fig. 5. As in Fig. 5, red solid curves are heading tuning.

The red dashed curves are the change in heading tuning (training minus pre-training). Blue curves are visual receptive fields. The blue dashed curve is the change in the visual receptive field (second probe minus first probe). Seven neurons from this dataset are also shown in Figs. 1, 2.



**Extended Data Fig. 9 | Controls for remapping experiments.** **a**, Data reproduced from Fig. 5e. Absolute change in visual receptive fields. Control flies navigated in a one-cue world (rather than a two-cue world) during the waiting period between the open-loop epochs used to compute the change in visual responses. In some cases (matched control), flies received exactly the same protocol as the experimental condition except with one-cue closed-loop epochs during the training period; in other words, these matched controls received 12 consecutive minutes of one-cue (rather than two-cue) closed-loop epochs during the training period. In all other cases (control), flies received 4-min blocks of one-cue closed-loop epochs interleaved with 150-s open-loop epochs during the training period, which lasted 12 min or more. **b**, Visual

receptive fields from control cells. Blue dashed curve is the change in visual receptive field (second probe minus first probe) over the control period. Typically, visual receptive fields were stable over time under control conditions (control neurons 2 and 3). On occasion, we observed spontaneous changes in the visual receptive field of an E-PG neuron during the control period (for example, control neuron 1), although these changes were not as large as the changes that we observed in many neurons in trained flies (see **a**). **c**, Heading tuning in the same three control cells. Note how the spontaneous changes in visual receptive fields seen in neuron 1 are accompanied by changes in heading tuning.

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | n/a                                 | Confirmed  |
|-------------------------------------|--|
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of all covariates tested   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. $F$ , $t$ , $r$ ) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted<br><i>Give <math>P</math> values as exact values whenever suitable.</i>                            |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's $d$ , Pearson's $r$ ), indicating how they were calculated  |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

### Software and code

Policy information about [availability of computer code](#)

Data collection	Matlab 2016a, Matlab 2017a, Matlab 2017b, ScanImage 2017, Fiji ( <a href="https://fiji.sc/">https://fiji.sc/</a> ), FicTrac ( <a href="http://rjdmoore.net/fictrac/">http://rjdmoore.net/fictrac/</a> )
Data analysis	All analyses of calcium imaging and electrophysiology data were performed using custom code written in Matlab 2017b (electrophysiology) and Matlab 2016b (calcium imaging) (see Methods for full description of analysis, code is available at <a href="https://github.com/wilson-lab/FisherLuDAlessandroWilson_AnalysisCode">https://github.com/wilson-lab/FisherLuDAlessandroWilson_AnalysisCode</a> ). Confocal images were analyzed using Fiji (ImageJ) and cell body counting was performed using the Fiji Cell Counter plugin (Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. Nat Methods 9, 676-682, (2012).)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For genetic perturbation experiments (Fig. 3c), the number of experiments performed was determined by first collecting a pilot data set (n=4 for each of the 3 genotypes using the R20A02-Gal4 driver line). Based on the initial effect size, power analysis was used to determine the number of experiments needed to test the hypothesis that visually-evoked inhibition was smaller in the experimental genotype. For all other experiments, sample sizes were chosen based on standard sample sizes in the field.
Data exclusions	<p>Figs. 1 &amp; 2: Epochs were included if the cell was healthy; specifically, this meant that the epoch-averaged voltage was below -33 mV and within 15 mV of the voltage observed at the start of the first epoch of the experiment, and also if the spike amplitude was &gt;50% of the amplitude observed in the first epoch. Closed-loop epochs were included if the fly visited all heading directions during that epoch. Cells were included if <math>\geq 2</math> open-loop epochs met these criteria; in Fig. 2 we also required that <math>\geq 2</math> closed-loop epochs met these criteria.</p> <p>Fig. 3: Cells were included if <math>\geq 2</math> open-loop epochs met our cell health criteria. A single recording from the UAS/+ control genotype was excluded because the biocytin fill showed that it was not an E-PG neuron.</p> <p>Fig. 4: 5/24 flies were excluded due to either weak fluorescence or an unstable offset between the angle of the E-PG bump and the fly's heading angle at the end of the initial closed-loop 1-cue epoch.</p> <p>Fig. 5: Cells were included if the epoch-averaged voltage from all epochs of the experiment (pre-training, training, post-training) was &lt;-33 mV, and if the fly visited all heading directions during the 2 epochs (8 min) of 1-cue closed-loop prior to training and during the final 2 epochs (8 min) of 2-cue closed-loop training. We required that the fly's mean yaw velocity was &gt;20°/s during the final 2 epochs of the 2-cue closed-loop training; 10 cells were excluded due to this restriction. We also removed recordings where the visual and/or heading turning curves were almost flat during the pre-training period (max-min <math>\leq 2</math>mV); 6 cells were removed due to this restriction.</p> <p>On occasion, during E-PG neuron electrophysiological recordings, we observed unexpected large inhibitory postsynaptic potentials with a stereotyped sharp onset, a large amplitude (&gt;15mV), and a stereotyped time course. They were followed by a prolonged period of depolarization when the variance of the voltage trace was also diminished. These events interfered with visual and heading tuning measurements, and so for Figs. 1-3, any epoch where such an event occurred was excluded from the analysis. For Fig. 5, the event was clipped but the rest of the epoch was used; 5% of open loop epochs and 10% of closed-loop epochs were clipped in this manner.</p>
Replication	For genetic perturbation experiments (Fig 3c) we reproduced the effect of R neuron silencing with two independent driver lines (R20A02 and R54E12). For all other experiments, results were replicated in different individuals within each data set.
Randomization	For genetic perturbation experiments (Fig 3c) flies were grouped based on genotype (neuronal activity manipulated by Kir2.1 expression vs control genotypes). For comparison between trained and control flies (Fig 5) the experimental protocol that would be performed (control vs training) was always decided on prior to starting the experiment. No other randomization was performed.
Blinding	The experimenter was not blind to genotype except when genetic perturbations were used: Figure 3c (Kir2.1 perturbation). For the Figure 3c data set collected for driver line R20A02-Gal4 the experimenter was blind to genotype after the pilot phase; because Fig. 3c pilot data were indistinguishable from subsequent data, all data were ultimately pooled, and overall the experimenter was blind to genotype in 67% of these recordings. For the data set obtained using the driver line R54E12-Gal4, the experimenter was not blind to genotype because the experimental genotype was obtained at a lower-than expected (sub-Mendelian) frequency, making it impractical to blind the experimenter.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Antibodies

Antibodies used	<p>-rat anti-FLAG (Novus Biologicals), Cat#: NBP1-06712B, RRID: AB_10006034, clone#: L5, lot#: B-3</p> <p>-rabbit anti-HA (Cell Signaling Technologies), Cat#: 3724 RRID: AB_1549585, clone#: C29F4, lot#: 9</p> <p>-DyLight 550-conjugated mouse anti-V5 (Bio-Rad), Cat#: MCA1360D550GA, RRID: AB_2687576, clone#: SV5-Pk1</p> <p>-mouse anti-Bruchpilot antibody (Developmental Studies Hybridoma Bank, nc82), Cat#: nc82, RRID: AB_2314866</p> <p>-chicken anti-GFP (Abcam), Cat#: ab13970, RRID: AB_300798</p>
Validation	<p>Multi-colorFlip Out (MCFO) immunohistochemistry:</p> <p>Multi-colorFlip Out (MCFO) genetic strategy (Nern et al. 2015) uses expression of epitopes (HA, FLAG and V5) that are not endogenous to the fly genome. For MCFO immunostaining in our study we followed the exact protocol as established and validated in Drosophila by Nern et al. that uses anti-HA, anti-FLAG and anti-V5 antibodies (Nern et al. 2015). These antibodies have also each been validated prior to Nern et al:</p> <p>rat anti-FLAG: Manufacturer notes confirms that rat anti-FLAG (Cat#: NBP1-06712B) has also been validated as FLAG-Tag specific in Drosophila (PMID: 26573957).</p> <p>rabbit anti-HA: Manufacturer confirmed rabbit anti-HA antibody has Epitope tag specificity using western blot and immunohistochemical analysis comparing untransfected with HA-tag transfected COS cells (<a href="https://www.cellsignal.com/products/primary-antibodies/ha-tag-c29f4-rabbit-mab/3724#validation-data">https://www.cellsignal.com/products/primary-antibodies/ha-tag-c29f4-rabbit-mab/3724#validation-data</a>).</p> <p>DyLight 550-conjugated mouse anti-V5: Manufacturer notes confirm that the DyLight 550-conjugated-Mouse anti V5-Tag, clone SV5-Pk1 recognizes the sequence, IPNPLLGLD, present on the P/V proteins of the paramyxovirus, SV5 (Dunn et al.1999) and can be used to detect recombinant proteins labeled with this V5-tag (Randall et al.1993 and Zhao et al. 2005).</p> <p>Other immunohistochemistry looking at neuron anatomy:</p> <p>The anti-Bruchpilot antibody (DSHB) is the standard in the Drosophila field as a background stain that labels presynaptic active zones to provide neurophil labeling for analysis of anatomy. This antibody was originally validated for use in Drosophila to label presynaptic active zones using immunohistochemistry and to be specific to Bruchpilot protein (Wagh et al. 2006).</p> <p>The anti-GFP antibody (Adcam) is the standard antibody used in the field for labeling exogenous expression of Green Fluorescent Protein (GFP) in Drosophila, note that this protein is not endogenously expressed in the Drosophila genome. Manufacturer's datasheet confirm that this anti-GFP antibody has been validated using western blot and immunohistochemistry to have specificity for Green Fluorescent Protein. Manufacturer also confirms the use of this antibody for immunolabeling of GFP in Drosophila across 121 peer-reviewer manuscripts (e.g. Sykes et al. 2005 PMID: 16122730).</p>

## Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	<p>Below is a complete description from the methods sections of all transgenic Drosophila strains use in this study included how they were obtained and their citation:</p> <p>The following Gal4 lines were obtained from the Bloomington Drosophila Stock Center (BDSC) and are described in ref. 40: P{R60D05-Gal4}attP2, P{R60D05-lexA}attP40, P{R19C08-lexA}attP40, P{R12B01-Gal4}attP2, P{R54E12-Gal4}attP2, P{R20A02-Gal4}attP2, P{GawB}EB1 was obtained from the BDSC and is described in Wang et al. 2002</p> <p>P{20XUAS-IVS-mCD8::GFP}attP40 was a gift from Barret Pfeiffer and Gerry Rubin and is described in Pfeiffer et al, 2010</p> <p>P{13XLexAop2-mCD8::GFP}attP40 was obtained from the BDSC and is described in Pfeiffer et al, 2010</p> <p>PBac{13xLexAop2-IVS-Syn21-Chrimson::tdT-3.1}VK00005 was a gift from Barret Pfeiffer and David Anderson and is described in Hoopfer et al. 2013</p> <p>P{20X-UAS-CsChrimson-tdTomato}VK00005 was a gift from John Tuthill who obtained it from Barret Pfeiffer. P{UAS-Hsap \KCNJ2.EGFP}7 was obtained from the BDSC and is described in Hardie et al. 2001</p> <p>P{UAS-GCamp6f}attP40 was obtained from the BDSC via Thomas Clandinin and is described in Chen et al. 2013</p> <p>Transgenes for MultiColor FlpOut were obtained from the BDSC and are described in Nern et al. 2015 these are w[1118] P{y[+t7.7] w[+mC]=GMR57C10-FLPG5.PEST}su(Hw)attP8; PBac{y[+mDint2], and w[+mC]=10xUAS(FRT.stop)myr::smGdP-HA}VK00005 P{y[+t7.7] , and w[+mC]=10xUAS(FRT.stop)myr::smGdP-V5-THS-10xUAS(FRT.stop)myr::smGdP-FLAG}su(Hw)attP1.</p>
Wild animals	No wild animals were used in this study.
Field-collected samples	No field samples were collected for this study.
Ethics oversight	No ethical approval was required because all experiments in this study were performed on Drosophila melanogaster.

Note that full information on the approval of the study protocol must also be provided in the manuscript.



# Generation of stable heading representations in diverse visual scenes

<https://doi.org/10.1038/s41586-019-1767-1>

Received: 29 December 2018

Accepted: 7 October 2019

Published online: 20 November 2019

Sung Soo Kim<sup>1,3,4\*</sup>, Ann M. Hermundstad<sup>1</sup>, Sandro Romani<sup>1</sup>, L. F. Abbott<sup>1,2</sup> & Vivek Jayaraman<sup>1\*</sup>

Many animals rely on an internal heading representation when navigating in varied environments<sup>1–10</sup>. How this representation is linked to the sensory cues that define different surroundings is unclear. In the fly brain, heading is represented by ‘compass’ neurons that innervate a ring-shaped structure known as the ellipsoid body<sup>3,11,12</sup>. Each compass neuron receives inputs from ‘ring’ neurons that are selective for particular visual features<sup>13–16</sup>; this combination provides an ideal substrate for the extraction of directional information from a visual scene. Here we combine two-photon calcium imaging and optogenetics in tethered flying flies with circuit modelling, and show how the correlated activity of compass and visual neurons drives plasticity<sup>17–22</sup>, which flexibly transforms two-dimensional visual cues into a stable heading representation. We also describe how this plasticity enables the fly to convert a partial heading representation, established from orienting within part of a novel setting, into a complete heading representation. Our results provide mechanistic insight into the memory-related computations that are essential for flexible navigation in varied surroundings.

Internal representations of the spatial relationship of an animal to its surroundings are essential for flexible navigation<sup>3,8–10</sup>. Although these representations must be stable to be useful for planning and goal-oriented behaviour, they must also adapt to changes in environmental and behavioural contexts. Indeed, the representations provided by head-direction cells, grid cells and place cells are all known to remap in different surroundings on the basis of spatially relevant sensory information<sup>23–26</sup>. A central question in navigation concerns how the brain carries out this flexible transformation of sensory information into a stable internal representation<sup>2,27</sup>. In insects, a multifunctional brain region known as the central complex<sup>11</sup> (Fig. 1a) has a key role in visually guided navigation, including flexible heading selection<sup>7,9,28</sup> and place learning<sup>29</sup>. Many of these abilities rely on successfully incorporating visual information from landscapes<sup>30</sup> or the pattern of polarized light and chromatic gradients in the sky<sup>4,5,31</sup> to generate an internal representation of heading in the central complex; specifically, a bump of activity in compass neurons (also known as E–PG neurons; see ‘Nomenclature’ in Methods) in the ellipsoid body<sup>3</sup>, a substructure of the central complex (Fig. 1a, b). These neurons are an important part of a ring attractor network<sup>32</sup> that maintains and updates the heading representation on the basis of self-motion<sup>33,34</sup> and visual signals<sup>3</sup>. Visual inputs are brought to the ellipsoid body by GABAergic ( $\gamma$ -aminobutyric-acid-releasing) ring neurons<sup>12</sup>, which have localized spatiotemporal receptive fields<sup>13–16</sup> (Fig. 1c). Here we show how network plasticity enables the flexible generation of a stable compass-neuron heading representation in different visual scenes.

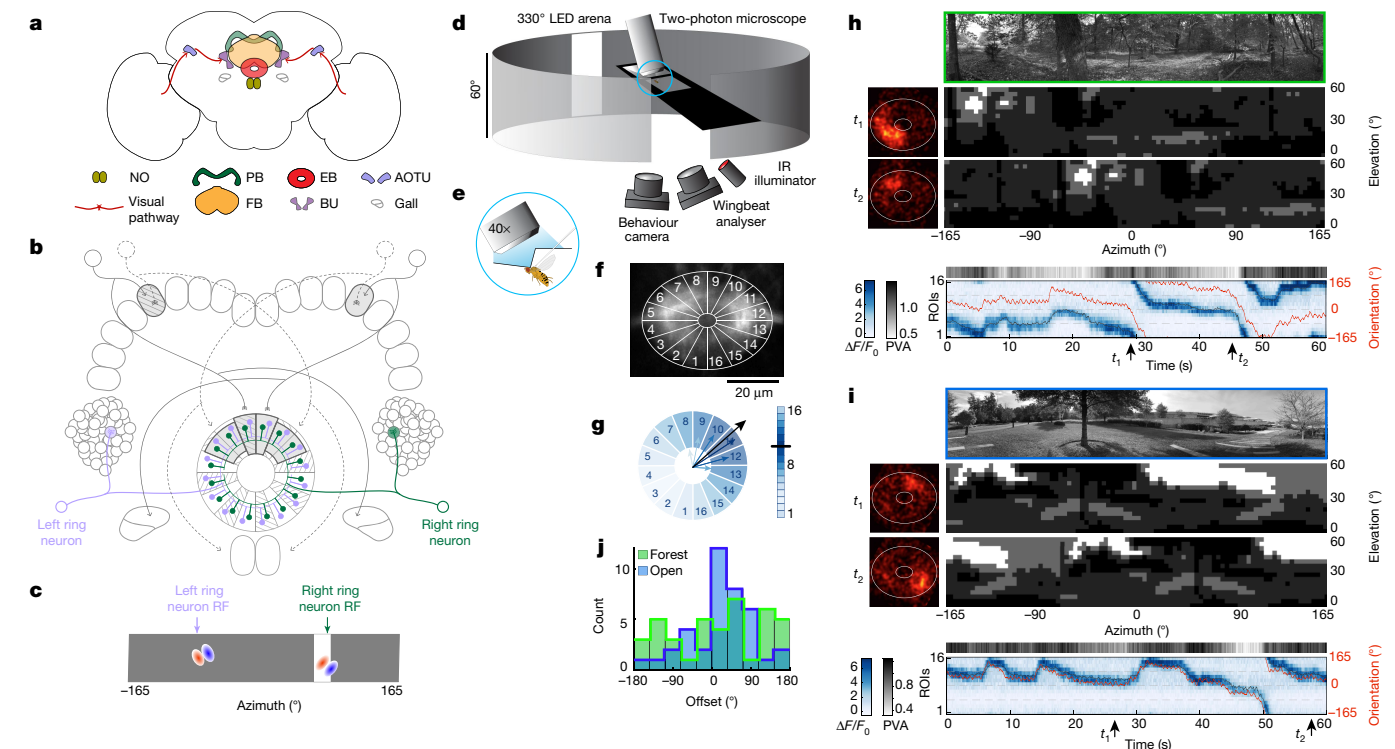
## Remapping of heading representation

To explore the flexibility of heading representation in the fly, we used two-photon calcium imaging to monitor responses of the

compass-neuron population in head-fixed flies that were flying in a virtual-reality arena, which consisted of panels of light-emitting diodes. The virtual-reality setup gave the insect one-dimensional, closed-loop control of its orientation<sup>32</sup> relative to visual scenes (Fig. 1d–g, Methods). Visual environments were derived from two natural scenes (Fig. 1h, i). The compass-neuron response in these scenes rapidly stabilized into an activity bump in the ellipsoid body that maintained a consistent angular relationship to the visual scene as the fly turned (Fig. 1h, i). Previous studies<sup>3,9,10,32</sup> in simpler visual settings (such as a single stripe) have shown that the bump tracks the visual scene, but with an offset between the angular position of the bump in the ellipsoid body and the angular orientation of the stripe relative to the fly. This pinning offset between the bump and visual cues (Methods) seldom changes across trials for a given fly in a specific visual setting, but differs across flies<sup>3,32–34</sup>. We found that the pinning offset also varied substantially across different naturally derived scenes for a single fly, and across flies for the same scene (Fig. 1j). We argue that this variable but stable offset is the natural outcome of plasticity in synapses that flexibly maps visual scenes onto the heading representation.

If activity-dependent plasticity between visual inputs and compass neurons underlies the observed variability in offset (Fig. 1j), experiencing an imposed artificial relationship between the scene and the bump should induce a sustained change in offset (as proposed for mammalian navigation systems<sup>17,18</sup>). A previous study of tethered flying flies used two-photon-localized optogenetics to temporarily displace a compass-neuron bump in the ellipsoid body by an arbitrary angle<sup>32</sup>. As in this previous study, here the original bump (Fig. 2a, d top) was quickly replaced by a displaced bump generated by focal optogenetics (Fig. 2b, Extended Data Fig. 1a). We then paired this artificial bump with an open-space scene (Fig. 1i) that was placed at a predetermined

<sup>1</sup>Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, VA, USA. <sup>2</sup>Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, USA. <sup>3</sup>Present address: Department of Molecular, Cellular, and Developmental Biology, University of California, Santa Barbara, Santa Barbara, CA, USA. <sup>4</sup>Present address: Neuroscience Research Institute, University of California, Santa Barbara, Santa Barbara, CA, USA. \*e-mail: sungsoo@ucsb.edu; vivek@janelia.hhmi.org



**Fig. 1 | E-PG neurons stably represent heading in different visual environments.** **a**, Central complex. Visual inputs to the ellipsoid body arrive from the optic lobe through the anterior optic tubercle to ring-neuron dendrites in the bulb<sup>14,15</sup>. EB, ellipsoid body; PB, protocerebral bridge; BU, bulb; FB, fan-shaped body; NO, noduli; AOTU, anterior optic tubercle. **b**, Ring neurons (purple and green) project from the bulb to the entire circumference of the ellipsoid body. E-PG or compass neurons (solid grey arrows) innervate single ellipsoid-body wedges. Circuit details have previously been published<sup>33</sup>. Dashed arrows, P-EN neurons (angular velocity). Small blobs in the ellipsoid body, synapses between ring and compass neurons. **c**, Fictive sample receptive fields (red, excitatory; blue, inhibitory) of two ring neurons (purple and green in **b**) shown in a flattened representation of the visual field (grey rectangle). The vertical stripe presented in the visual arena activates the green ring neuron. RF, receptive field. **d**, Imaging setup. IR, infrared; LED, light-emitting diode. **e**, Tethered flying fly. **f**, Ellipsoid body segmented into 16 regions of

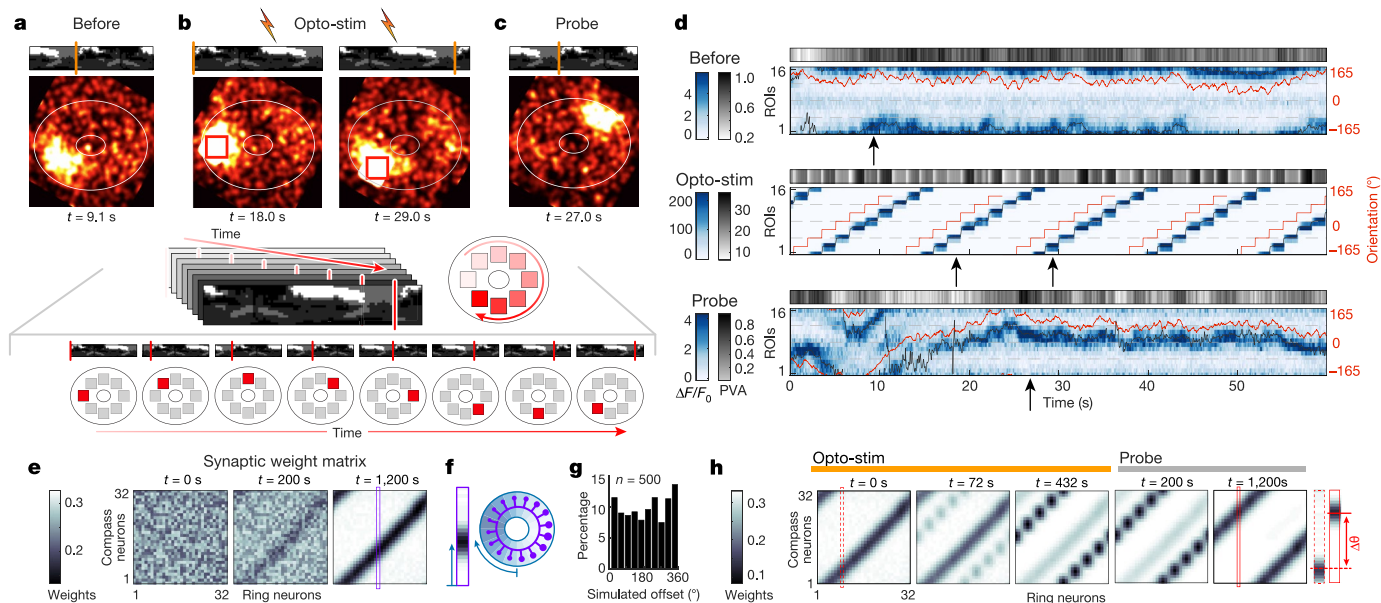
interest (ROIs). **g**, Population vector average (PVA) of  $\Delta F/F_0$  computed to obtain the angular position and amplitude of the compass-neuron activity bump. **h**, Compass-neuron calcium transients during closed-loop tethered flight in a visual environment derived from a natural scene (forest; shown at the top). Middle, actual scene presented on an arena of blue light-emitting diodes with discretized brightness. Snapshots of compass-neuron activity in the ellipsoid body at times  $t_1$  and  $t_2$ , corresponding to different scene orientations. Bottom,  $\Delta F/F_0$  of 16 ROIs over time. Greyscale band, PVA amplitude. Red line, scene orientation. GCaMP signal colour-coded in blue. Black line, PVA. **i**, Calcium transients from the fly in **h** in a different scene (top), open space. **j**, Distribution of mean pinning offset across flies. Offset distribution for the open-space scene is significantly different from uniform for unknown reasons (open-space scene, 39 trials, 10 flies, unimodality test by randomization,  $P < 0.0001$ ; forest scene, 40 trials, 10 flies,  $P = 0.3603$ ).

angular position in the arena relative to the bump (Fig. 2b, d middle, Supplementary Video 1). We repeatedly shifted the artificial bump through eight positions around the ellipsoid body, while simultaneously shifting the scene around the visual arena to maintain its fixed angular relationship to the imposed bump (Fig. 2b, d middle). A 5-min pairing protocol was sufficient to change the offset, and the newly imposed relationship between the visual scene and the compass-neuron bump was clearly preserved in subsequent closed-loop probe trials (Fig. 2c, d bottom, Extended Data Fig. 1f, h, i). Such remapping could also be induced with simpler visual scenes (such as a single stripe; Extended Data Fig. 1b, e, g, i, m), but could not be induced without the optogenetic reagent or in darkness (Extended Data Fig. 1j, k, n, o). Thus, we find strong experimental support for plasticity that enables visual surroundings to be flexibly remapped onto the compass-neuron population upon sustained experience of a specific angular relationship between the bump and the scene.

## Plasticity creates a stable compass

The experience-dependent remapping that we observed (Fig. 2a–d), which involves co-activation of specific visual inputs and compass neurons, is strongly suggestive of Hebbian plasticity, which has been hypothesized to explain how mammalian head-direction cells tether

to visual cues<sup>17,18</sup>. We built an anatomically motivated circuit model to better understand the effect of such a plasticity mechanism on scene-to-bump remapping. The key components of the model (Fig. 2e–h; implementation details are given in the Supplementary Information) are: (i) visual ring neurons that distribute information about visual features to all compass neurons throughout the ellipsoid body<sup>13–15,35</sup> (Fig. 1b, c, 2f)—for simplicity, we treat ring-neuron receptive fields as encoding only azimuthal information, and address the two-dimensional spatiotemporal complexity of their responses<sup>14</sup> in a later section; (ii) ring attractor dynamics, a form of all-to-all competitive network dynamics that ensures a single compass-neuron bump that can remain active in darkness<sup>32–34</sup>; (iii) a plasticity rule through which the co-activation of GABAergic inhibitory ring neurons and compass neurons results in a depression of the synaptic weight between them<sup>36</sup> (inhibitory Hebbian plasticity<sup>17–21</sup>), whereas the activation of compass neurons alone results in potentiation (alternative plasticity rules are given in Supplementary Information). In this model (which shares some conceptual similarities with recent models of mammalian head-direction cells<sup>20</sup> and grid cells<sup>22</sup>), the turns that the fly undertakes cause a retinotopic shift of the visual stimulus (which activates a different set of ring neurons), and angular velocity signals that are carried by so-called P-EN neurons<sup>33,34</sup> (dotted lines in Fig. 1b) rotate the compass-neuron bump. For a stable heading representation, bump positions driven by visual input and



**Fig. 2 | Manipulation of heading representation pinning offset.** **a–d**, Activity snapshots of compass neurons before (**a**), during (**b**) and after (**c**) optogenetic manipulation in an open loop (imposed natural-scene orientations at the top, with vertical red lines emphasizing the relative orientations). Extended Data Figure 1a provides details of the optogenetic stimulation (opto-stim) protocol. **a**, Original pinning offset (arrow in **d**, top, shows the time of this snapshot). **b**, Optogenetic imposition of new offset. **b**, Top left, bump imposed on left side of the ellipsoid body (below, red rectangle) when scene oriented as at the top. **b**, Top right, 45° counter-clockwise rotated scene and bump with offset as in left. **b**, Middle, sequence of optogenetically imposed ellipsoid-body offsets (**d**, middle) (Methods). **b**, Bottom, expanded view of same sequence as shown in **b** (middle). **c**, After manipulation. The bump position relative to the same visual scene orientation as in **a**, shifted by offset imposed in **b** (compare **d**, top and bottom). **d**, Compass neuron activity before (top), during (middle) and after (bottom) optogenetic manipulation (Supplementary Video 1). Arrow in the top panel corresponds to **a**; arrows in the middle panel correspond to the left and

right panels of **b** (top); and arrow in the bottom panel corresponds to **c**. **e**, Simulation snapshots. Time-varying synaptic weights between ring and compass neurons (Extended Data Fig. 2). Simulation begins with random synaptic weights (left). Synapses between coactive ring and compass neurons are weakened. Synapses from inactive ring to active compass neurons are potentiated (see Supplementary Information for different plasticity rules). The weight matrix stabilizes over time (right) (Supplementary Video 2). Vertical purple rectangle, sample mapping from ring neuron 16 to all compass neurons. **f**, Simulated compass neurons when ring neuron 16 is active. **g**, Distribution of bump offsets across 500 simulations. **h**, Simulated optogenetic bump shift. Left, weight matrix before manipulation. Second and third from the left, a new map develops while the existing map weakens. Rightmost two panels, consolidation of the new map during a probe trial. Dashed red rectangle, initial synaptic weights from ring neuron 9 to compass neurons. Solid red rectangle, same weights after consolidation; offset shifted.

angular velocity should be in register. That is, for any given heading, plasticity should ensure that inhibitory ring neurons create a position of decreased inhibition in the ellipsoid body that coincides with where the P–EN input moves the bump—essentially a self-consistent mapping of visual cues onto the bump.

We first tested the model for a simple scene with a single vertical stripe (Extended Data Fig. 1b–e), simulating the fly turning through the scene (Fig. 2e–g, Supplementary Video 2; a complex scene is shown in Extended Data Fig. 2a–c). These rotations ensured both that the bump travelled around the ellipsoid body and that ring neurons corresponding to all visual-feature positions were selectively co-activated at appropriate angular orientations. Starting with random synaptic weights, Hebbian plasticity produced a spatially consistent mapping and stable offset between the heading representation and the angular position of the single visual feature (Fig. 2e). Simulating optogenetic manipulation as a current injection into model compass neurons reproduced the remapping phenomenon (Fig. 2h, Extended Data Fig. 2d, e). These results account for the varying offsets observed across flies<sup>3</sup>, the persistence of an offset for a given scene in a single fly and the flexibility that allows the ellipsoid body to track heading within different visual scenes.

### Optogenetic inversion of the map

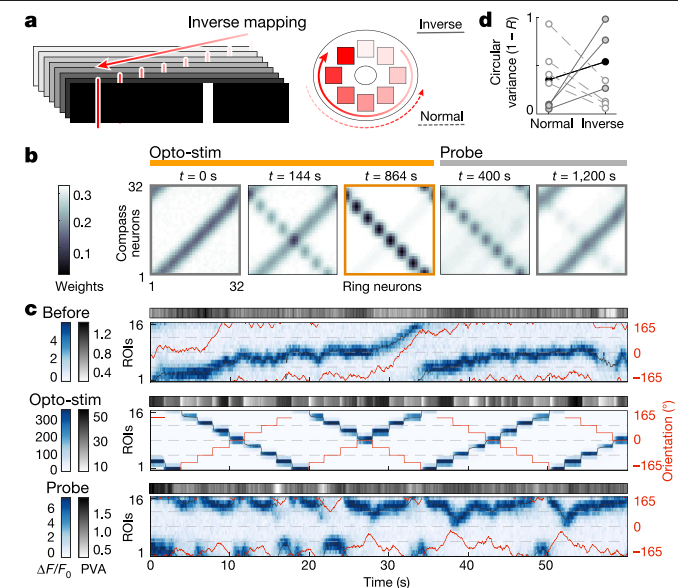
In further simulations, the natural concurrence between scene movement and bump position during turns could be inverted, with visual

cues overriding self-motion input to drive the bump backwards (Fig. 3a, b). In optogenetic offset-induction experiments, we found that the actual network was indeed flexible enough to induce an inverted remapping in which visual input drove the bump around the ellipsoid body in the opposite direction than would be expected (Fig. 3c, d, Supplementary Video 3). In the model, the inversion was eventually corrected after prolonged ring attractor dynamics driven by self-motion (Fig. 3b rightmost panel), but the short trial duration in our physiological experiments probably limited our ability to observe such a correction in vivo. Thus, although self-motion exerts a strong influence over bump movement, network plasticity allows for a strong and notably flexible driving role for visual cues.

### Remapping after experiencing ambiguity

Ring attractor dynamics ensures a single heading representation at any given time even for complex scenes, but under some circumstances this can be unstable<sup>4</sup>. For example, a scene with two identical stripes at diagonally opposite locations (Extended Data Fig. 3a) makes orientation within the scene inherently ambiguous<sup>3</sup>. Our model predicts that, upon prolonged exposure to this two-stripe scene, the plasticity mechanism creates a visual map with two potential offset angles. If a single-stripe scene is then presented, this results in two competing heading representations, with the ring attractor network selecting one of them at any particular time (Extended Data Fig. 3b, c). We found a similar effect experimentally in some probe trials after just 5 min of in vivo



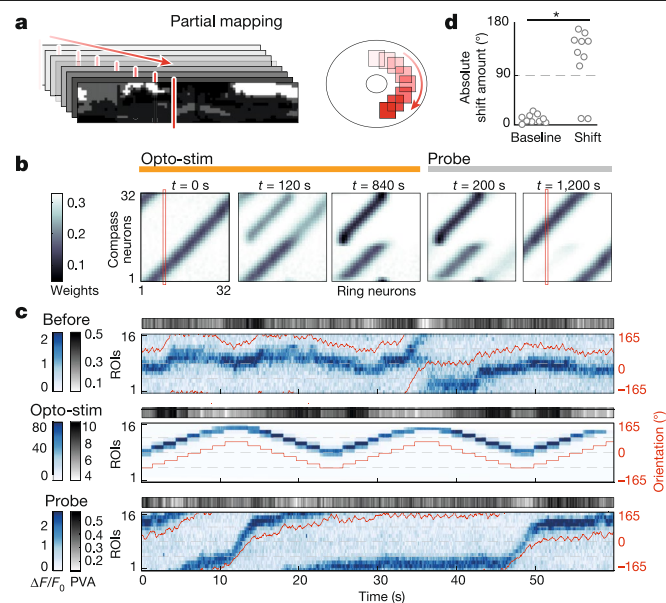


**Fig. 3 | Optogenetically imposed inverse mapping of visual scene onto compass neurons.** **a**, Inverse mapping protocol, in which the stripe is angularly displaced opposite to the optogenetic bump displacement. **b**, Simulation of inverse mapping. Inverse mapping is complete after 864 s, and maintained during initial period of probe trial (left panel under ‘probe’). Sustained angular velocity input eventually corrects the map in simulations (right panel under ‘probe’). **c**, Segments (60 s) of in vivo calcium transients before (top), during (middle) and after (bottom) a 10-min manipulation. Before manipulation, the bump followed the direction of stripe motion (top). After manipulation, bump motion mirrors stripe motion but in the opposite angular direction (bottom) (Supplementary Video 3). **d**, Circular variance of bump offset during the probe trial, computed for the normal arrangement of ellipsoid-body ROIs (normal), and for the inverse arrangement of ellipsoid-body ROIs (inverse). Four out of eight flies tested showed a smaller circular variance for the inverted arrangement of ellipsoid-body ROIs (white dots), indicating that the map was indeed inverted. Poor bump tracking—resulting from incomplete map manipulation—was observed in one fly, resulting in intermediate circular variances for both maps (black solid dots). Grey solid dots, three flies maintained the correct map.

closed-loop experience with a two-stripe scene in the absence of any optogenetic manipulation (Extended Data Fig. 3d–i, Supplementary Video 4). A companion study<sup>37</sup> to this Article finds electrophysiological and imaging signatures of offset switches in a larger fraction of experiments after walking flies experience such ambiguous scenes for longer durations. These results demonstrate how exposure to an ambiguous visual scene can, through the interactive influence of plasticity and ring attractor dynamics, affect the reliability of an otherwise-stable heading representation.

### Building a full map from partial views

In our remapping experiments thus far, the fly performed multiple complete rotations to establish a stable heading representation in a novel setting, which seems unlikely under natural conditions. *Drosophila* can see nearly 320° of the visual scene from a single orientation<sup>38</sup> and the E-PG bump typically activates more than 90° of the ellipsoid body<sup>3</sup>; this suggests that even limited experience of a scene should trigger Hebbian plasticity that affects a large sector of the ellipsoid body. In the model, we found that full mapping of a visual scene could occur even if the bump was rotated only by 180° or less during optogenetic manipulation (Fig. 4a, b, Extended Data Fig. 4). We directly tested this prediction by imposing an angular relationship between a vertical stripe and an artificial compass-neuron bump, but this time limiting the

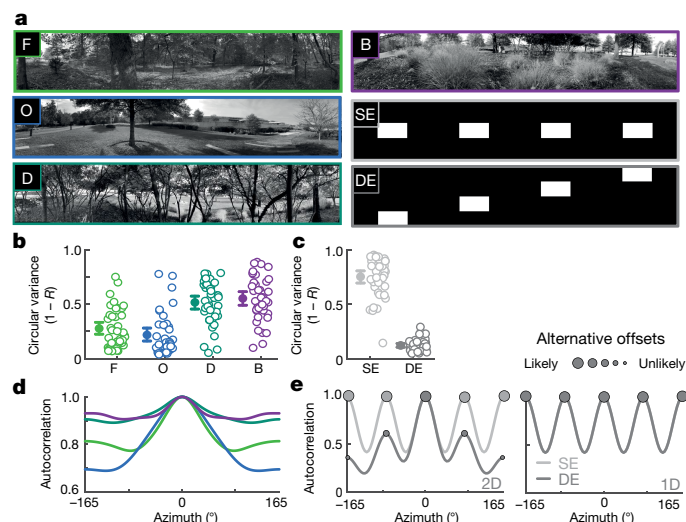


**Fig. 4 | Experience of only 180° of rotation during optogenetic manipulation suffices to induce global remapping.** **a**, Experimental protocol in which optogenetic manipulation and the experience of scene orientations span only 180°. **b**, Simulation of protocol with a simple single-stripe scene. After manipulation ( $t = 840$  s under optogenetic stimulation), there are two sets of weak synapses (top left and bottom left), and the upper right corner of the weight matrix is completely erased. During the probe trial, a newly imposed offset propagated across the entire weight matrix (probe). **c**, Segments (60 s) of compass-population calcium transients before (top), during (middle) and after (bottom) optogenetic manipulation, spanning 180° of the ellipsoid body and using a naturalistic scene as oriented in **a**. Compare the offsets in the top and bottom panels. **d**, Distribution of absolute offset shift across flies. Left, baseline before manipulation. Right, offset shift by manipulation (10 flies, two-sided bootstrap test of mean difference,  $*P = 0.0002$ ).

range of bump positions to 180°. Indeed, we found that—in the majority of flies (6 out of 10)—experiencing this limited range of bump positions was sufficient to induce a stable heading that matched the imposed offset in the probe period of the trial (Extended Data Fig. 4d, e). We could successfully induce a full remapping of the single-stripe scene in a few flies even in a more-constrained situation in which the range of bump positions spanned only 60° (in 7 out of 20 flies) (Extended Data Fig. 4i–k). Further analysis revealed that successful remapping was more likely when the stripe and the bump started inside the newly mapped region in the probe trial, consistent with simulations (Extended Data Fig. 4f–h, j, k). This probably occurred because the internally generated angular velocity signal could move the bump into regions that were not previously traversed while still preserving the new offset, thereby allowing the new heading representation to stabilize. We also observed full remapping after limited-angle exposure in experiments with a natural scene (Fig. 4a, c, d). These results provide insights into how Hebbian plasticity combined with ring attractor dynamics enables the fly to convert information gathered from limited views of a novel scene into a complete heading representation within that scene.

### Stability of the compass in two-dimensional scenes

Looking across all experiments, we observed that heading representations exhibit a varying degree of stability across different scenes (Fig. 5a, b). We wondered whether structure in the vertical dimension—typical for natural scenes and known to be encoded by visual ring neurons<sup>13,14,39</sup>—could resolve potential ambiguities in scenes with repeating visual features in the horizontal dimension (for example,



**Fig. 5 | The stability of bump dynamics is predicted by two-dimensional information in visual scenes.** **a**, Four natural scenes (F, forest (Fig. 1h); O, open field (Fig. 1i); D, dense forest; and B, bush) were downsampled and discretized. Two artificial scenes with the same local features at the same (SE) and different elevations (DE) were also used. **b**, Circular variance of instantaneous pinning offset with natural scenes (4 repetitions of 2 scenes per fly, 40 trials from 10 flies for each condition) (Methods). The bump reliably tracked the orientation of forest and open-space scenes (indicated by low circular variance). F, mean = 0.2771, 95% confidence interval = [0.2231, 0.3344]; O, mean = 0.2180, 95% confidence interval = [0.1616, 0.2828]. Tracking was poor (a high circular variance) for dense-forest and bush scenes. D, mean = 0.5163, 95% confidence interval = [0.4557, 0.5752]; B, mean = 0.5528, 95% confidence interval = [0.4893, 0.6135]. Bootstrap tests of the difference in mean circular variance between each pair of scenes showed significant difference across all pairs (two-sided,  $P < 0.0001$ ), except between forest and open-space scenes (two-sided,  $P = 0.169$ ) and between dense-forest and bush scenes ( $P = 0.406$ ). **c**, Circular variance of instantaneous pinning offset for the same-elevation and different-elevation artificial scenes (4 repetitions of both scenes per fly, 40 trials from 10 flies) (Methods). The offset was stable (a low circular variance) for the different-elevation scene (mean = 0.1212, 95% confidence interval = [0.1046, 0.1392]), but not for the same-elevation scene (mean = 0.7521, 95% confidence interval = [0.6937, 0.8045]). The mean circular variance between scenes was significantly different (two-sided bootstrap test,  $P < 0.0001$ ). **d**, Two-dimensional autocorrelation of natural scenes. **e**, One-dimensional (1D; right) and two-dimensional (2D; left) autocorrelation of the same-elevation and different-elevation artificial scenes; the one-dimensional autocorrelation is identical, but the two-dimensional autocorrelations are different.

the ‘same-elevation’ scene in Fig. 5a). Using artificial stimuli, we found that the bump reliably tracked the orientation of an artificial scene with four identical objects placed at different elevations, whereas it could not stably track when these objects were placed at the same elevation (Fig. 5c). This stability is well-predicted by the fact that the two-dimensional autocorrelation of each scene is distinctly single-peaked (Fig. 5d, e). We conclude that the two-dimensional organization of a scene<sup>13,14,39</sup> contributes to the generation and stability of the pinning offset.

Some insects are capable of snapshot-based navigation<sup>30,31,40,41</sup> in which stored visual scenes are recalled to drive scene-specific directional actions. Further analysis of our model indicated that multiple visual maps can be stored simultaneously if plasticity between visual ring neurons and compass neurons is presynaptically gated and the network has access to a rich ring-neuron representation of visual scenes<sup>15,35</sup> (Extended Data Figs. 5, 6, Supplementary Information). Other spatially informative sensory inputs—including spectral<sup>42</sup>, mechanical (for example, wind<sup>43</sup>) and olfactory cues<sup>44</sup>—may also contribute to differentiating natural sensory environments.

## Discussion

We have shown how inhibitory Hebbian plasticity can rapidly transform visual feature information into an attractor-driven internal representation. Angular velocity input to the attractor converts an emerging mapping on the basis of limited views of a scene into a complete and consistent heading representation, a potentially critical function in animal navigation. The induction of inverse maps emphasizes the notable flexibility of the system. A key issue that remains unresolved is the nature of bump dynamics during translation in a two-dimensional environment. Mammalian head-direction cells are unaffected by translation<sup>1</sup>, but our model suggests that the compass circuit tracks the angle between the orientation of the fly and an object in the visual scene without correcting for translation—potentially making it a local compass. However, the plasticity that we have identified required only a few minutes, and may be even faster under natural conditions when the system can co-opt an existing mapping from ring to compass neurons. In our simulations (data not shown), this timescale prevented nearby objects and transient stimuli—such as neighbouring conspecifics that would not move coherently with the bearing of the fly—from being mapped, but tethered the compass to distant objects that moved coherently with the turns of the fly.

The locus of plasticity is likely to be synapses between ring and compass neurons; this idea is also favoured by the authors of the accompanying Article<sup>37</sup>, who present electrophysiological evidence that is consistent with plasticity altering inhibitory visual inputs to individual compass neurons. At a synaptic and biophysical level, it remains to be seen how the Hebbian mechanism that we have proposed relates to, and interacts with, other forms of plasticity such as spike-timing-dependent plasticity<sup>45,46</sup>, or with plasticity-inducing mechanisms such as nitric oxide signalling in the ellipsoid body<sup>47</sup>, dopaminergic modulation (as seen in the fly mushroom body<sup>36,48</sup>) or plateau potentials (as seen during remapping of hippocampal place cells<sup>49</sup>).

Our results support a model in which plasticity is constantly active to allow rapid adaptation to new settings, enabling the ring attractor to generate a single heading direction even in a complex environment. Such stable sensorimotor representations probably enable animals to overcome transient uncertainties in their surroundings as they pursue diverse behavioural goals.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1767-1>.

1. Taube, J. S. The head direction signal: origins and sensory-motor integration. *Annu. Rev. Neurosci.* **30**, 181–207 (2007).
2. Geva-Sagiv, M., Las, L., Yovel, Y. & Ulanovsky, N. Spatial cognition in bats and rats: from sensory acquisition to multiscale maps and navigation. *Nat. Rev. Neurosci.* **16**, 94–108 (2015).
3. Seelig, J. D. & Jayaraman, V. Neural dynamics for landmark orientation and angular path integration. *Nature* **521**, 186–191 (2015).
4. Heinze, S. & Reppert, S. M. Sun compass integration of skylight cues in migratory monarch butterflies. *Neuron* **69**, 345–358 (2011).
5. Heinze, S. & Homberg, U. Maplike representation of celestial E-vector orientations in the brain of an insect. *Science* **315**, 995–997 (2007).
6. Varga, A. G. & Ritzmann, R. E. Cellular basis of head direction and contextual cues in the insect brain. *Curr. Biol.* **26**, 1816–1828 (2016).
7. el Jundi, B. et al. Neural coding underlying the cue preference for celestial orientation. *Proc. Natl Acad. Sci. USA* **112**, 11395–11400 (2015).
8. Butler, W. N., Smith, K. S., van der Meer, M. A. A. & Taube, J. S. The head-direction signal plays a functional role as a neural compass during navigation. *Curr. Biol.* **27**, 1259–1267 (2017).
9. Giraldo, Y. M. et al. Sun navigation requires compass neurons in *Drosophila*. *Curr. Biol.* **28**, 2845–2852 (2018).



10. Green, J., Vijayan, V., Mussells Pires, P., Adachi, A. & Maimon, G. A neural heading estimate is compared with an internal goal to guide oriented navigation. *Nat. Neurosci.* **22**, 1460–1468 (2019).
11. Turner-Evans, D. B. & Jayaraman, V. The insect central complex. *Curr. Biol.* **26**, R453–R457 (2016).
12. Hanesch, U., Fischbach, K. F. & Heisenberg, M. Neuronal architecture of the central complex in *Drosophila melanogaster*. *Cell Tissue Res.* **257**, 343–366 (1989).
13. Seelig, J. D. & Jayaraman, V. Feature detection and orientation tuning in the *Drosophila* central complex. *Nature* **503**, 262–266 (2013).
14. Sun, Y. et al. Neural signatures of dynamic stimulus selection in *Drosophila*. *Nat. Neurosci.* **20**, 1104–1113 (2017).
15. Omoto, J. J. et al. Visual input to the *Drosophila* central complex by developmentally and functionally distinct neuronal populations. *Curr. Biol.* **27**, 1098–1110 (2017).
16. Shiozaki, H. M. & Kazama, H. Parallel encoding of recent visual experience and self-motion during navigation in *Drosophila*. *Nat. Neurosci.* **20**, 1395–1403 (2017).
17. Skaggs, W. E., Knierim, J. J., Kudrimoti, H. S. & McNaughton, B. L. A model of the neural basis of the rat's sense of direction. *Adv. Neural Inf. Process. Syst.* **7**, 173–180 (1995).
18. Knierim, J. J. in *Head Direction Cells and the Neural Mechanisms of Spatial Orientation* (eds Wiener S. I. & Taube J. S.) 163–185 (MIT Press, 2005).
19. Cope, A. J., Sabo, C., Vasilaki, E., Barron, A. B. & Marshall, J. A. A computational model of the integration of landmarks and motion in the insect central complex. *PLoS ONE* **12**, e0172325 (2017).
20. Page, H. J. I. & Jeffery, K. J. Landmark-based updating of the head direction system by retrosplenial cortex: a computational model. *Front. Cell. Neurosci.* **12**, 191 (2018).
21. Campbell, M. G. et al. Principles governing the integration of landmark and self-motion cues in entorhinal cortical codes for navigation. *Nat. Neurosci.* **21**, 1096–1106 (2018).
22. Ocko, S. A., Hardcastle, K., Giocomo, L. M. & Ganguli, S. Emergent elasticity in the neural code for space. *Proc. Natl Acad. Sci. USA* **115**, E11798–E11806 (2018).
23. Knierim, J. J., Kudrimoti, H. S. & McNaughton, B. L. Interactions between idiothetic cues and external landmarks in the control of place cells and head direction cells. *J. Neurophysiol.* **80**, 425–446 (1998).
24. Fyhn, M., Hafting, T., Treves, A., Moser, M. B. & Moser, E. I. Hippocampal remapping and grid realignment in entorhinal cortex. *Nature* **446**, 190–194 (2007).
25. Solstad, T., Boccara, C. N., Kropff, E., Moser, M. B. & Moser, E. I. Representation of geometric borders in the entorhinal cortex. *Science* **322**, 1865–1868 (2008).
26. Krupic, J., Bauza, M., Burton, S., Barry, C. & O'Keefe, J. Grid cell symmetry is shaped by environmental geometry. *Nature* **518**, 232–235 (2015).
27. Connor, C. E. & Knierim, J. J. Integration of objects and space in perception and memory. *Nat. Neurosci.* **20**, 1493–1503 (2017).
28. Neuser, K., Triphan, T., Mronz, M., Poeck, B. & Strauss, R. Analysis of a spatial orientation memory in *Drosophila*. *Nature* **453**, 1244–1247 (2008).
29. Ofstad, T. A., Zuker, C. S. & Reiser, M. B. Visual place learning in *Drosophila melanogaster*. *Nature* **474**, 204–207 (2011).
30. Collett, T. S. & Zeil, J. Insect learning flights and walks. *Curr. Biol.* **28**, R984–R988 (2018).
31. el Jundi, B. et al. A snapshot-based mechanism for celestial orientation. *Curr. Biol.* **26**, 1456–1462 (2016).
32. Kim, S. S., Rouault, H., Druckmann, S. & Jayaraman, V. Ring attractor dynamics in the *Drosophila* central brain. *Science* **356**, 849–853 (2017).
33. Turner-Evans, D. et al. Angular velocity integration in a fly heading circuit. *eLife* **6**, e23496 (2017).
34. Green, J. et al. A neural circuit architecture for angular integration in *Drosophila*. *Nature* **546**, 101–106 (2017).
35. Omoto, J. J. et al. Neuronal constituents and putative interactions within the *Drosophila* ellipsoid body neuropil. **12**, 103 (2018).
36. Hattori, D. et al. Representations of novelty and familiarity in a mushroom body compartment. *Cell* **169**, 956–969 (2017).
37. Fisher, Y. E., Lu, J., D'Alessandro, I. & Wilson, R. I. Sensorimotor experience remaps visual input to a heading-direction network. *Nature* <https://doi.org/10.1038/s41586-019-1772-4> (2019).
38. Buchner, E. *Dunkelanregung des Stationaeren Flugs der Fruchtfliege Drosophila*. Dipl. thesis, Univ. Tübingen (1971).
39. Dewar, A. D. M., Wystrach, A., Philippides, A. & Graham, P. Neural coding in the visual system of *Drosophila melanogaster*: how do small neural populations support visually guided behaviours? *PLOS Comput. Biol.* **13**, e1005735 (2017).
40. Judd, S. P. D. & Collett, T. S. Multiple stored views and landmark guidance in ants. *Nature* **392**, 710–714 (1998).
41. Narendra, A., Gourmaud, S. & Zeil, J. Mapping the navigational knowledge of individually foraging ants, *Myrmecia croslandi*. *Proc. R. Soc. Lond. B* **280**, 20130683 (2013).
42. Longden, K. D. Colour vision: a fresh view of lateral inhibition in *Drosophila*. *Curr. Biol.* **28**, R308–R311 (2018).
43. Suver, M. P. et al. Encoding of wind direction by central neurons in *Drosophila*. *Neuron* **102**, 828–842 (2019).
44. Jacob, P. Y. et al. An independent, landmark-dominated head-direction signal in dysgranular retrosplenial cortex. *Nat. Neurosci.* **20**, 173–175 (2017).
45. Song, S., Miller, K. D. & Abbott, L. F. Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. *Nat. Neurosci.* **3**, 919–926 (2000).
46. Cassenaer, S. & Laurent, G. Hebbian STDP in mushroom bodies facilitates the synchronous flow of olfactory information in locusts. *Nature* **448**, 709–713 (2007).
47. Kuntz, S., Poeck, B. & Strauss, R. Visual working memory requires permissive and instructive NO/cGMP signaling at presynapses in the *Drosophila* central brain. *Curr. Biol.* **27**, 613–623 (2017).
48. Aso, Y. & Rubin, G. M. Dopaminergic neurons write and update memories with cell-type-specific rules. *eLife* **5**, e16135 (2016).
49. Bittner, K. C. et al. Conjunctive input processing drives feature selectivity in hippocampal CA1 neurons. *Nat. Neurosci.* **18**, 1133–1142 (2015).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

# Article

## Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

### Nomenclature

We follow an abbreviation convention that is agreed upon by most research groups working on the central complex<sup>50</sup>. For E-PG or compass neurons<sup>32</sup>, E (ellipsoid body) before a ‘-’ represents predominantly spiny and putatively postsynaptic processes, and P (protocerebral bridge) and G (gall) after a ‘-’ represent predominantly bouton-like and probable presynaptic processes. When fully expanded, the abbreviation E-PG stands for PB<sub>G1-8</sub>.b-EBw.s-D/Vgall.b<sup>50</sup>. Similarly, P-EN neurons (Fig. 1), which arborize in the noduli (N), refer to PB<sub>G2-9</sub>.s-EBt.b-NO1.b neurons<sup>50</sup>.

### Terminology

In the manuscript, we use the term heading representation to describe what it is that the E-PG neurons encode. However, the representation often persists when a tethered fly is standing still on a ball<sup>3</sup>—that is, when it has no heading in a strict sense. On the basis of such data, we would define heading as the angular orientation of the body axis of the fly in a visual scene. Future experiments may well determine that E-PG neurons represent the head direction of the fly, but all E-PG imaging experiments thus far—including those in this study—have been performed on head-fixed flies in tethered preparations, leaving this issue unresolved.

### Fly stocks

Fly stocks have previously been described<sup>32,33</sup>. In brief, flies with either a codon-optimized UAS-GCaMP6<sup>F1</sup> or a recombinant of UAS-CsChrimson-mCherry-tag<sup>52</sup> and UAS-GCaMP6<sup>F1</sup> or codon-optimized UAS-GCaMP6<sup>F1</sup> were driven by split-GAL4<sup>53,54</sup> SS00096 from the Rubin laboratory. All experiments were performed with 6–10-day-old female flies. Flies were randomly picked from their housing vials for all experiments. All flies were raised from the egg stage on standard cornmeal and soybean-based medium<sup>55</sup> or with additional 0.2 mM all-*trans*-retinal<sup>52</sup> for flies with CsChrimson.

### Fly preparation for imaging during head-fixed flight

The procedure for fly preparation has previously been described<sup>32</sup>. In brief, flies were anaesthetized on a cold plate at 4 °C. The front legs were removed and the proboscis was pressed into its head capsule and immobilized with wax to minimize brain movement. The fly was tethered at the tip of a tungsten wire and positioned under a custom-designed stainless-steel shim, as previously described<sup>13,56,57</sup>. The back of the head capsule was kept nearly vertical to maximize exposure of the eyes of the fly to the surrounding light-emitting-diode (LED) arena. UV-curable adhesive was used to fix the head under the shim, then the cuticle at the top of the head and fat cells were carefully removed and trachea were carefully pushed to the back of the brain to optically reveal the central brain.

### Visual stimulation

**Visual arena.** The hardware has been described<sup>32</sup>. In brief, a female fly was placed at the centre of the arena and visual stimuli were presented on a vertically placed cylindrical LED display<sup>58</sup> spanning 330° in azimuth and 60° in elevation. The display was covered with multiple layers of colour filter to avoid excessive leak into a photon detector and a diffuser to avoid reflection<sup>3,13,56</sup>. The wingbeat amplitude of each wing was computed online by analysing images acquired with a camera, using custom-built image analysis software written in MATLAB, similar to a previously described method<sup>57</sup>. The image acquisition rate of the camera was 119.2 Hz, which was slow enough to capture the full

shadow of wings to compute the wingbeat amplitude. For closed-loop experiments, the gain was 5.1° per second for each degree of the difference between the left and right wingbeat amplitudes ( $\Delta$ WBA)<sup>59</sup>. Air was manually puffed at the fly if it stopped flying. The data during this stalled period were excluded from analyses.

**Stimuli.** We used various visual stimuli. Natural scenes were derived from panoramic photographs taken at the Janelia Research Campus. Using the full luminance resolution of the arena resulted in excessive leak into a photon detector even after multiple layers of filters, making it impossible to detect bump position (especially with extremely low laser power used for simultaneous imaging and optogenetic stimulation). Further, the level of light at full luminance was enough to activate CsChrimson in most flies. To reduce the light leak and undesired activation of CsChrimson, we downsampled and monochromatized natural scene photographs (Fig. 1h, i, 4a, 5a) to four luminance levels close to a log scale (0, 2, 6 and 15). Other visual stimuli included a bright vertical stripe spanning 60° in elevation and 15° in azimuth (Fig. 3a, Extended Data Figs. 1b, 3b, 4d, i), two bright vertical stripes 165° apart (Extended Data Fig. 3a), a random dot pattern of which each pixel is either of maximum brightness or dark, and patterns containing four small horizontal bars each spanning 30° in azimuth and 15° in elevation (Extended Data Fig. 5a, b). All the stimuli used in this study were presented on a blue LED arena. We used a greyscale in the figures for visual clarity. To avoid a sudden luminance change that might induce a startle response in flies, the 30° arena gap behind the fly was stitched in all protocols to maintain overall luminance. Thus, when an object crosses the gap, it does not disappear but jumps across it.

### Protocols

**Optogenetic bump offset shift.** An experiment (14 flies) (Extended Data Fig. 1b–d, i, m) began with a 1-min exposure to a closed-loop random dot stimulus (trial 1). It was followed by 3 1-min closed-loop single-stripe trials (trials 2–4), a 5-min optogenetic manipulation trial that imposes a fixed 90° offset between the bump and a scene (trial 5), 3 1-min closed-loop single-stripe trials (trials 6–8), another 5-min optogenetic trial with –90° offset (trial 9), and 2 1-min closed-loop single-stripe trials (trials 10 and 11). Each trial was followed by a 15-s dark trial before the next trial started. During optogenetic manipulation trials, 8 positions in the ellipsoid body, separated by 45° (with a visual stimulus of a corresponding offset), were sequentially stimulated, each of which took approximately 2–2.5 s. The initial position of the visual stimulus during closed-loop trials was random. Trial 2 was used for flies to establish a stable offset. Trials 3 and 4 and trials 7 and 8 were used to measure the baseline variability of the bump offset within a single fly before optogenetic manipulation. Trials 6 and 7 and trials 10 and 11 were used to measure the baseline variability after optogenetic manipulation. Trials 4 and 6 were used to measure the effect of optogenetic manipulation in trial 5 (90° offset). Trials 8 and 10 were used to measure the effect of optogenetic manipulation in trial 9 (–90° offset). Control experiments (ten flies each) used the same order of trials except that either CsChrimson was not expressed (Extended Data Fig. 1j, n) or the stripe was not presented (Extended Data Fig. 1k, o) during manipulation trials. A natural scene was also tested (Fig. 2a–d, Extended Data Fig. 1f, h, l). To increase statistical power, all data collected before or after the –90° protocol were rotated 180° and pooled with the 90° protocol during analyses.

**Bump offset shift with two vertical stripes.** The order of trials was identical to optogenetic bump offset shift experiments, but—during manipulation trials—two stripes at opposite sides of the visual field (165° apart in the 330° arena) were presented under closed-loop control (Extended Data Fig. 3d–i). Trials 6 and 10 were used to measure the number of bumps and the bump offset variance for the initial 15 s after manipulation trials, and trials 7 and 11 were used as control trials. Ten flies were tested.

**Forced optogenetic inverse mapping.** There were two 1-min single-stripe closed-loop trials followed by 10 min of an optogenetic inverse mapping trial and 2 min of a probe trial (Fig. 3). Consecutive trials were separated by a 3-s dark trial.

**Natural scene protocols.** Two 2-min closed-loop trials with a down-sampled and monochromatized forest scene were presented (trials 1 and 2). They were followed by 22-min closed-loop trials with an open-space scene (trials 3 and 4), and all 4 trials were repeated (trials 5–8). All consecutive trials were separated by a 5-s dark trial. The initial scene orientation of each trial was random. Trials 2 and 5 were used to measure the offset shift between two forest-scene trials separated by open-space scene trials. Trials 4 and 7 were used to measure the offset shift between two open-space scene trials separated by forest-scene trials. Trials 2 and 3 were used to measure the offset shift during the transition from a forest scene to an open-space scene. Trials 4 and 5 were used to measure the offset shift during the transition from an open-space scene to a forest scene. Ten flies were tested (Fig. 1h, i, 5b, d). The whole protocol was repeated for another pair of less-reliable natural scenes (dense forest and bush) (Fig. 5a, b, d). Finally, to address the relevance of two-dimensional organization of the visual scene to the bump position computation, the same protocol was repeated with 2 scenes of 4 artificial objects: in each scene, four horizontal objects were presented with equal azimuthal separation and either the same or different elevations (Fig. 5a, c, e).

**Bump offset shift with limited optogenetic manipulation.** An experiment (Fig. 4, Extended Data Fig. 4) began with a 1-min closed-loop trial with a single stripe (trial 1). It was followed by a 2-min closed-loop single-stripe trial (trial 2), a 30-s open-loop probe trial (trial 3), a 5-min open-loop manipulation trial (trial 4), a 30-s open-loop probe trial (trial 5), a 2-min closed-loop trial (trial 6), a 30-s open-loop probe trial (trial 7), a 5-min open-loop manipulation trial (trial 8), a 30-s open-loop probe trial (trial 9) and a 2-min closed-loop trial (trial 10). All consecutive trials (except the probe trials following manipulation trials) were separated by a 3-s dark trial. The initial scene orientation of closed-loop trials was random. During trial 2, the bump offset was roughly determined by visual inspection. Then, a target offset was determined to be 180° away from this baseline offset and optogenetically imposed during manipulation trials. Three manipulation protocols were used (ten flies each). The first protocol (local protocol 1) spanned 60° of the ellipsoid body, in which 3 positions separated by 30° were optogenetically stimulated. Each position was stimulated for 1.5–2.5 s in sequence. The probe trials were composed of the same visual stimuli used during optogenetics trials to measure the effectiveness of the optogenetic manipulation. The position of a stripe in closed-loop probe trials began at the middle of the range of stripe positions used during manipulation. The second protocol (local protocol 2) spanned 60° of the ellipsoid body, in which three positions separated by 30° were optogenetically stimulated. Each position was stimulated for 1.5–2.5 s in sequence. During probe trials, two stripe positions (one at the centre of the manipulated area and another 180° away from it) were repeatedly presented (each for 3 s) to probe the global effect of local manipulations. The position of a stripe in closed-loop probe trials was random. For further analysis, flies from the two protocols (local protocols 1 and 2) were pooled (Extended Data Fig. 4j, k) and regrouped depending on the position of the bump and the stripe at the beginning of the probe trial. The final protocol (local protocol 3) spanned 180° of the ellipsoid body (Extended Data Fig. 4d), in which 8 positions separated by 22.5° were optogenetically stimulated. The same probe stimuli as in local protocol 2 were used in addition to 8 stripe positions separated by 45° to cover all orientations. The offset during probe trials was measured over the final 5 s. The final protocol was repeated with a natural scene (Fig. 4).

The position of the pattern, wingbeat amplitudes, air-puffing signal and two-photon frame trigger were all simultaneously collected using custom software written in MATLAB that used National Instrument data acquisition hardware.

### Two-photon calcium imaging

Calcium imaging was performed using a custom-built two-photon microscope<sup>60</sup>. We used a 40× objective (NA 1.0, 2.8 mm WD) and a GaAsP photomultiplier tube (PMT). A Chameleon Ultra II laser tuned to 930 nm with a custom-built pulse compressor was used as the excitation source with a maximum power of 8 mW at the sample. We used the same saline as in previous studies<sup>3</sup> with adjusted calcium concentration at 2.0 mM. We imaged the ellipsoid body over 6-plane volumes using a fast remote focusing technique<sup>61</sup>, which was modified in-house, at a rate of 9.8 Hz volume rate (256 × 256 resolution, 58.8-Hz frame rate) with an equal spacing of 3–6 μm between individual scanning planes. The objective was tilted by 30° to enable imaging of the ellipsoid body with the head of the fly at a natural, vertical angle.

### Two-photon optogenetic stimulation

The protocol used was largely along previously described lines<sup>32</sup>, but differed in a few details. A single two-photon laser source was used for both imaging and optogenetic stimulation, by temporally modulating the laser power, which was implemented using the PowerBox feature in ScanImage<sup>60</sup> replacing the custom MATLAB software described in previous work<sup>32</sup> (Extended Data Fig. 1a). Increased two-photon efficiency owing to a pulse compressor allowed a lower laser power for imaging and optogenetic stimulation than previously described<sup>32</sup>. For the calcium-imaging-only period, a maximum laser power of 2 mW was used for both forward and backward scanning phases. During optogenetic stimulation of CsChrimson, the laser power was kept the same except for the defined stimulation area only during the forward scanning phase, in which a maximum laser power of 30 mW (typically 20 mW) was used. To prevent tissue damage, this laser power was manually adjusted during each trial to a minimal power that was sufficient to develop a bump at the site of stimulation. On average, the optogenetically induced GCaMP signal measured during the backward scanning phase was 13.3% greater than the normal condition across flies (one-tailed paired *t*-test, *P* = 0.022) in the optogenetic bump-shifting experiment with a natural scene. This higher-than-natural activity was required to inhibit the naturally generated bump. However, two vertical-stripe protocol results indicate that plasticity can be induced at the natural activity level.

### Data analysis

We used MATLAB for data analysis. To avoid bias, no statistical methods were used to predetermine the power and the sample size. The fixed-offset optogenetic experiment used 14 flies, and the forced optogenetic inverse mapping experiments relied on 8 flies. All other experiments were performed until data from 10 flies were collected.

**Calculation of fluorescence changes.** The background noise level was predetermined by measuring the oscillatory noise from the PMT. This level was then subtracted from all imaging data, and the data were half-rectified before further analysis. A running average intensity projection of a volume (six planes) at a given time was computed for each pixel. Then, 16 ROIs were manually assigned, as previously described<sup>32</sup>. Next, time series for each ROI were obtained by taking the average of the fluorescence signal within the ROI at each point in time. For calcium imaging experiments without optogenetics,  $\Delta F/F_0$  was computed using  $F_0$  as the mean of the lowest 10% of signals in each ROI. No further temporal smoothing was applied.

**PVA of a bump and its amplitude.** As a simple measure of the bump position and strength, the PVA was computed as the weighted vector

# Article

average across ellipsoid-body wedges, with the weight determined by the fluorescence level ( $\Delta F/F_0$ ), and the vector determined by the position of each ROI in the ellipsoid body. The amplitude of the PVA was determined as the length of the average vector. We used brewermap (S. Cobeldick, MathWorks file exchange) with a colour scheme 'blue' from <http://colorbrewer2.org/> to depict all PVA plots.

**Calculation of the number of bumps.** For each frame, a bump was defined as any contiguous set of ROIs with  $\Delta F/F_0$  greater than a threshold value (defined in each frame to be the mean  $\Delta F/F_0$  across ROIs + 1 s.d.)<sup>3</sup> (Extended Data Fig. 3h).

**Offset between the estimated bump position and the pattern position, and offset deviation.** For a given trial, the first 15 s were discarded, as were time points when the fly did not fly, which were determined by the wingbeat amplitude. The offset between the absolute scene orientation (to the experimenter) and the PVA estimate was calculated as the mean angular difference for the remaining time. The deviation was calculated as the circular variance. The visual arena (covering 330°) was mapped to 360°, as was the position of the scene.

**Analysis of optogenetic offset manipulation trials.** The exact artificial offset imposed by optogenetic stimulation during manipulation trials was determined by the mean angular difference between the scene orientation and the PVA during optogenetic stimulation.

**Circular linearity test.** For the optogenetic manipulation protocol, the expected amount of offset shift was assumed to be the same as the artificially imposed amount of shift. The sum of absolute angular difference between these two values across flies was used as a test statistic. To obtain the null distribution, the observed amounts of shift were randomized across flies and the sum of absolute angular differences was calculated, all of which was repeated 10,000 times. The *P* value was calculated by counting the number of outcomes from randomization that were smaller than the test statistic (Extended Data Fig. 1h–k).

**Circular unimodality or circular asymmetry test.** We used this test to determine whether a set of directional data was significantly unimodal or asymmetric. The circular variance of the data was used as a test statistic. Each data point was assigned a random direction sampled from a circularly uniform distribution, after which the circular variance was calculated. This random assignment procedure was repeated 10,000 times to generate a null distribution. The *P* value was determined by the number of times at which the circular variance was smaller than the test statistic (Fig. 1j, Extended Data Fig. 5c). This method reliably works only for unimodal data and may generate false-negative results for multimodal data.

**Bootstrap test of the mean difference.** This test was used to establish the difference of means of two datasets when they did not satisfy the assumption of Gaussian distributions. The difference of means of two datasets was used as a test statistic. Two sets of data were pooled, random samples were assigned to each group either with (bootstrap) or without (randomization) replacement, and the difference of the means of the two groups were calculated. This process was repeated 10,000 times to generate the null distribution. The *P* value was computed by counting the number of events with an outcome that was greater than the test statistic (Extended Data Fig. 1l–o). Random sampling both with and without replacement generated similar *P* values in all tests in our study.

**Circular variance of pinning offset.** The variance in pinning offset relative to each scene (Fig. 5b, c) was computed as the circular variance of the instantaneous pinning offset along the time of a single trial. Each fly experienced four repetitions of two scenes. For each scene, all trials were pooled across flies (in total, 40 trials each).

**Circular variance of inverse map.** The circular variance of the bump offset during the probe trial was calculated for both normally arranged ellipsoid-body ROIs and inversely arranged ellipsoid-body ROIs. If the circular variance of the latter was smaller than the former, the mapping from the visual scene orientation to compass neurons was determined to be inverted (Fig. 3d).

**Binomial exact test.** For Extended Data Fig. 4j, the baseline probability of flies shifting their offsets by more than 90° is 1 out of 7 if the stripe starts outside manipulated positions (red dots). Assuming binomial sampling from this distribution, the chance of 6 or more flies out of 13 shifting their offsets by more than 90° (blue dots) is  $P = 0.0059$ . For Extended Data Fig. 4k, the baseline probability of flies shifting their offsets by more than 90° is 3 out of 16 if the stripe or bump starts outside the manipulated positions (red dots). The chance of all 4 flies shifting their offsets by more than 90° (blue dots) assuming binomial sampling with a probability of 3/16 is  $P = 0.0012$ .

**Natural scene analysis.** Each scene was smoothed with a two-dimensional Gaussian filter with a s.d. of 4 pixels (Extended Data Fig. 5e). Then, the two-dimensional autocorrelation of each scene was calculated (Fig. 5d). Each scene was tiled horizontally (three copies) and the top and the bottom were padded with zeros. Then, MATLAB function `xcorr2` was applied to this tiled scene, and to another scene representing the centre of this tiled scene. The middle range of azimuth values of the outcome (corresponding to the azimuthal range of one scene within the tiled image) was finally normalized by the maximum value to obtain two-dimensional autocorrelation. The one-dimensional autocorrelation was obtained by first taking the average intensity of the smoothed scene over elevation, then applying `xcorr` between this one-dimensional trace and a concatenated version of this trace, and finally normalizing by the maximum value. The two-dimensional cross-correlation was computed in the same way, except that `xcorr2` was applied to two tiled scenes: one scene with three horizontal copies of itself padded at the top and bottom, and another scene without horizontal copies but padded at the top and bottom.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

All data are freely available at [http://research.janelia.org/jayaraman/Kim\\_etal\\_Nature2019\\_Downloads/](http://research.janelia.org/jayaraman/Kim_etal_Nature2019_Downloads/).

## Code availability

All code is freely available at [http://research.janelia.org/jayaraman/Kim\\_etal\\_Nature2019\\_Downloads/](http://research.janelia.org/jayaraman/Kim_etal_Nature2019_Downloads/).

- Wolff, T. & Rubin, G. M. Neuroarchitecture of the *Drosophila* central complex: a catalog of nodulus and asymmetrical body neurons and a revision of the protocerebral bridge catalog. *J. Comp. Neurol.* **526**, 2585–2611 (2018).
- Chen, T. W. et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
- Klapoetke, N. C. et al. Independent optical excitation of distinct neural populations. *Nat. Methods* **11**, 338–346 (2014).
- Luan, H., Peabody, N. C., Vinson, C. R. & White, B. H. Refined spatial manipulation of neuronal function by combinatorial restriction of transgene expression. *Neuron* **52**, 425–436 (2006).
- Pfeiffer, B. D. et al. Refinement of tools for targeted gene expression in *Drosophila*. *Genetics* **186**, 735–755 (2010).
- Guo, A. et al. Conditioned visual flight orientation in *Drosophila*: dependence on age, practice, and diet. *Learn. Mem.* **3**, 49–59 (1996).
- Seelig, J. D. et al. Two-photon calcium imaging from head-fixed *Drosophila* during optomotor walking behavior. *Nat. Methods* **7**, 535–540 (2010).
- Maimon, G., Straw, A. D. & Dickinson, M. H. Active flight increases the gain of visual motion processing in *Drosophila*. *Nat. Neurosci.* **13**, 393–399 (2010).

58. Reiser, M. B. & Dickinson, M. H. A modular display system for insect behavioral neuroscience. *J. Neurosci. Methods* **167**, 127–139 (2008).
59. Bahl, A., Ammer, G., Schilling, T. & Borst, A. Object tracking in motion-blind flies. *Nat. Neurosci.* **16**, 730–738 (2013).
60. Pologruto, T. A., Sabatini, B. L. & Svoboda, K. ScanImage: flexible software for operating laser scanning microscopes. *Biomed. Eng. Online* **2**, 13 (2003).
61. Rupprecht, P., Prendergast, A., Wyart, C. & Friedrich, R. W. Remote z-scanning with a macroscopic voice coil motor for fast 3D multiphoton laser scanning microscopy. *Biomed. Opt. Express* **7**, 1656–1671 (2016).

**Acknowledgements** We thank A. Jenett, T. Wolff and G. Rubin for sharing the split line SS00096; B. Pfeiffer, A. Wong, D. Anderson and G. Rubin for sharing codon-optimized GCaMP6f DNA; C. Dan for codon-optimized GCaMP6f flies; Janelia Fly Core and, in particular K. Hibbard and S. Coffman, for fly husbandry; J. Liu for virtual-reality support; V. Goncharov and C. McRaven for microscope support; J. Arnold for fly holder design; Vidrio for ScanImage support; T. Kawase for animation; and E. Nielson and S. Houck for operational support. We are grateful to A. Karpova and members of V.J.'s and A.M.H.'s laboratories for useful discussions and comments on the manuscript. S.S.K., A.M.H., S.R. and V.J. are supported by Howard

Hughes Medical Institute; L.F.A. is supported by NSF NeuroNex Award DBI-1707398, the Gatsby Charitable Foundation and the Simons Collaboration for the Global Brain.

**Author contributions** S.S.K., A.M.H., L.F.A. and V.J. conceptualized the project; S.S.K. undertook all experiments; S.S.K. performed modelling, in collaboration with L.F.A., A.M.H. and S.R.; S.S.K., A.M.H. and V.J. provided the visualizations; S.S.K. and V.J. wrote the initial draft, and all authors contributed to editing.

**Competing interests** The authors declare no competing interests.

#### **Additional information**

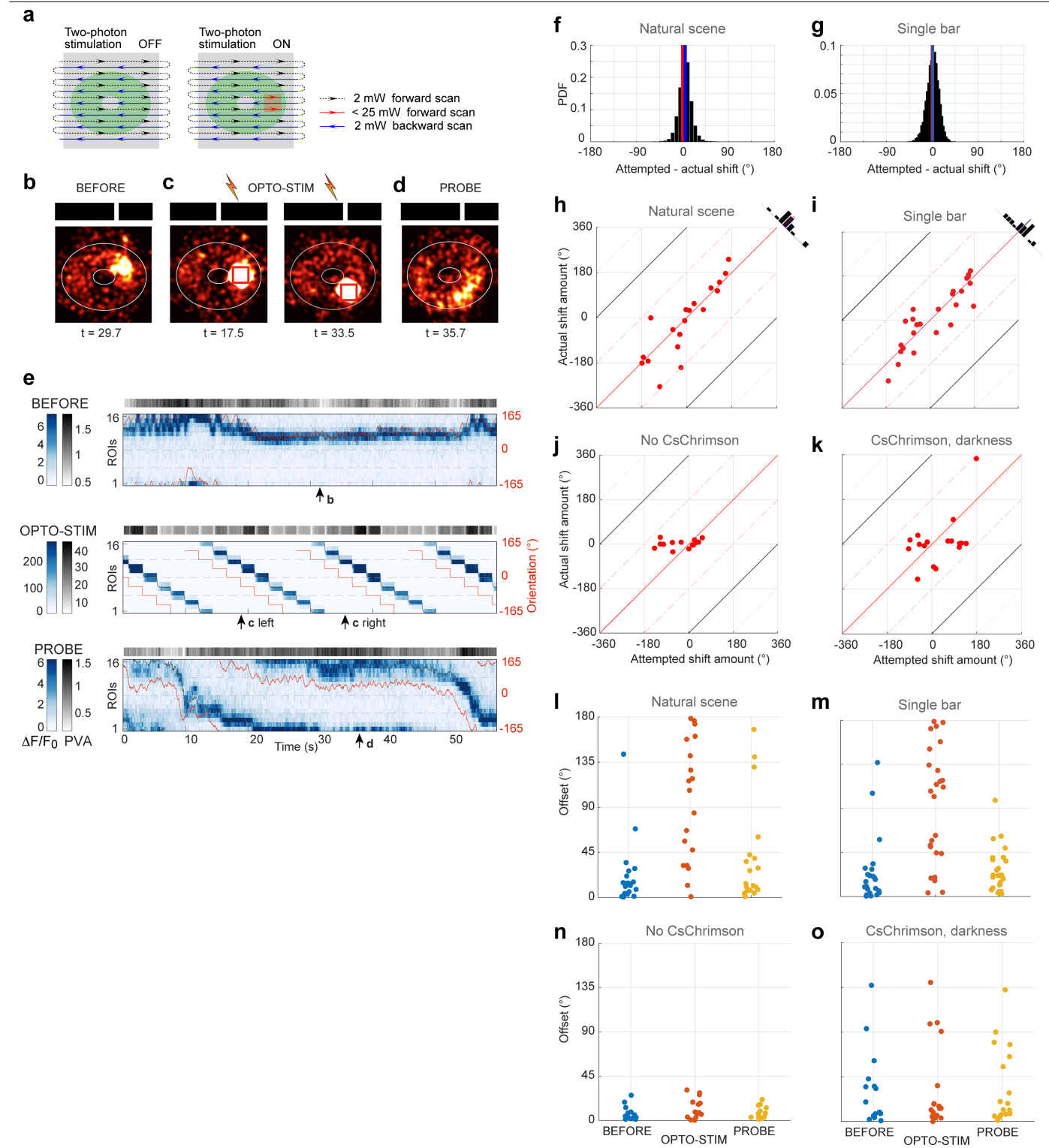
**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1767-1>.

**Correspondence and requests for materials** should be addressed to S.S.K. or V.J.

**Peer review information** *Nature* thanks Holger G. Krapp and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



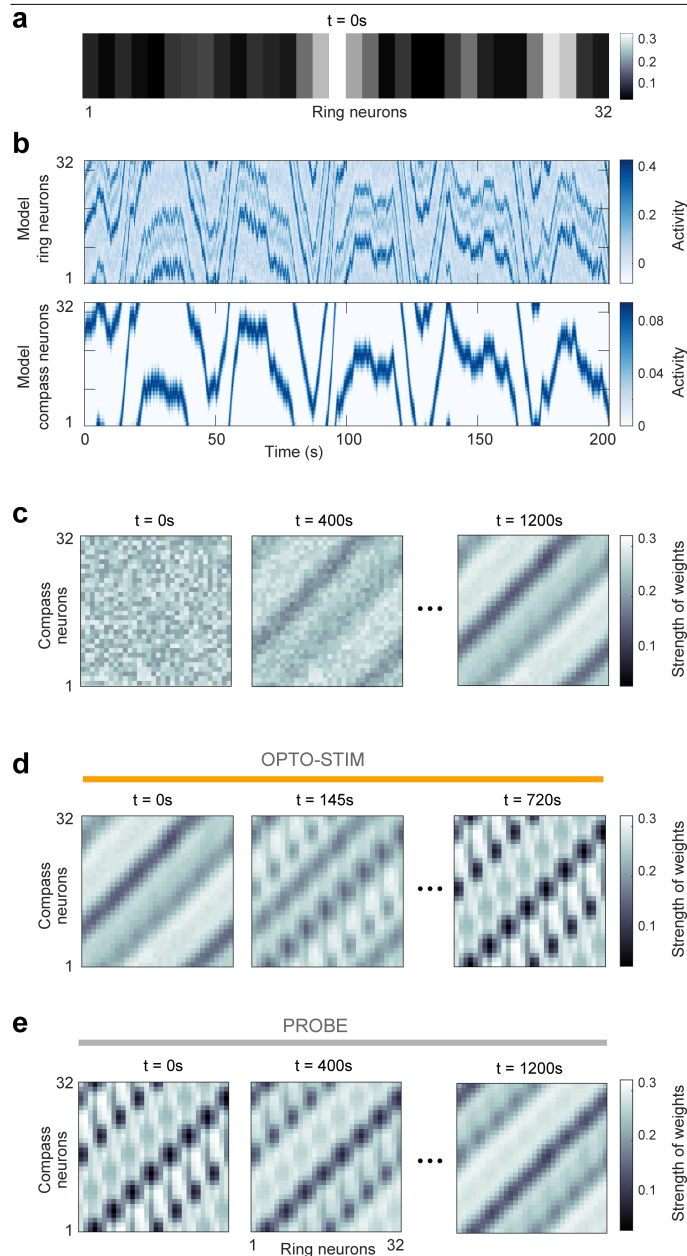


**Extended Data Fig. 1** | See next page for caption.

**Extended Data Fig. 1 | Manipulation of pinning offset of heading representation relative to visual scene.**

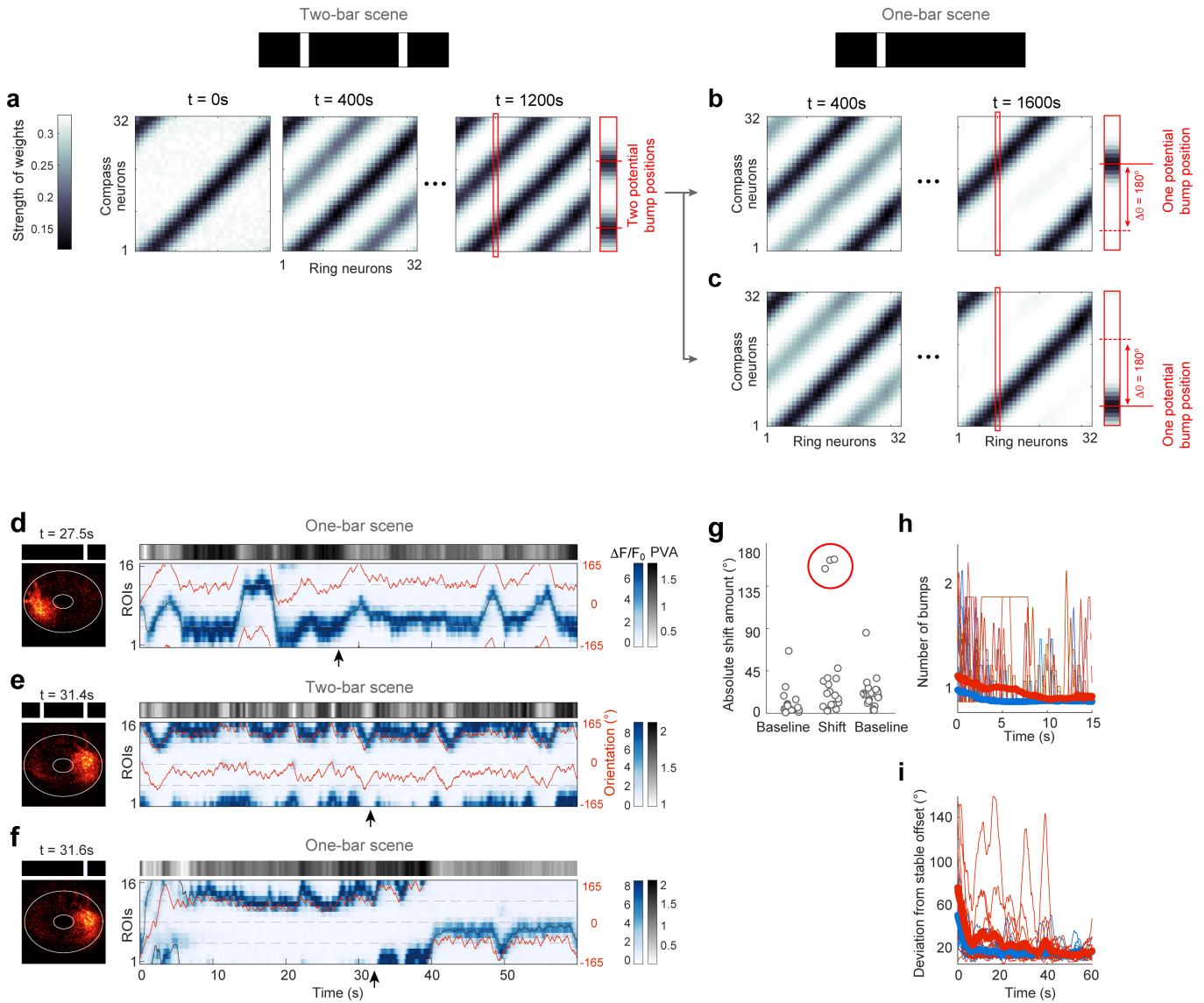
**a**, Schematic shows simultaneous calcium imaging and localized optogenetic stimulation. **b–d**, Snapshots of compass-neuron population activity before, during and after optogenetic manipulation in open loop (orientations of imposed single-stripe visual scene are shown at the top). **b**, A bump offset of close to zero before optogenetic manipulation (arrow in **e** shows the time of this snapshot). **c**, Optogenetic imposition of the new offset. Left, when the vertical stripe is in front of the fly, the bump was imposed on the right side of the ellipsoid body (rectangle). Right, 45° rotated scene and bump with the same offset as shown on the left. This offset was sequentially imposed across eight positions of the visual scene and ellipsoid body for approximately 2 s per position for 5 min (**e** middle). **d**, Snapshot of compass-neuron calcium transients after manipulation (**e** bottom). The bump position relative to same visual scene as in **b** is now shifted by the offset imposed in **c**. **e**, Segments (60 s) of imaging before (top), during (middle) and after (bottom) a 5-min optogenetic manipulation. Conventions are the same as in Fig. 1. **f**, Bootstrapped distribution of the mean difference between the imposed and actual offset shifts in Fig. 2 (natural scene), which was not significantly different from 0 (19 trials from 10 flies, bootstrapped mean difference test, two-sided,  $P = 0.6276$ ). **g**, Bootstrapped distribution of the mean difference between the imposed and actual offset shifts in **b–d** (single stripe), which was not significantly different from 0 (25 trials from 14 flies, two-sided,  $P = 0.8932$ ). **h–k**, Distribution of imposed

(x axis) versus actual (y axis) offset shifts across flies. The distribution is significantly linear along the identity line (circular linearity test. **h**, Natural scene, 19 trials from 10 flies,  $P < 0.0001$ . **i**, Single stripe, 25 trials from 14 flies,  $P < 0.0001$ . **j**, No CsChrimson, 14 trials from 10 flies,  $P = 0.0934$ . **k**, In darkness, 17 trials from 10 flies,  $P = 0.6064$ ). **l–o**, Absolute change in offset across two trials before manipulation (blue) and across two trials after manipulation (yellow), and absolute change in offset induced by manipulation (red). Bootstrapped mean difference tests, one-sided.  $n$  values are the same as in **h–k**. **l**, Natural scene, bootstrapped mean difference test between epochs before and during manipulation,  $P = 0.0464$ ; and between epochs during and after manipulation,  $P = 0.0024$ . **m**, Single stripe, bootstrap tests of the mean difference showed a significant difference between the baseline offset shifts and manipulated offset shifts ( $P = 0.0207$  between epochs before and during manipulation; and  $P = 0.0252$  between epochs during and after manipulation). **n**, No CsChrimson control, bootstrap tests of the mean difference did not show any significant difference;  $P > 0.05$  for all pairs. **o**, Darkness control, bootstrap tests of the mean difference did not show any significant difference;  $P > 0.05$  for all pairs. Baseline offset shifts were comparable to the experimental group (**m**), but greater than the control group without CsChrimson (**n**). This suggests that the baseline offset variance in the experimental group might be due to a higher baseline activity of the compass-neuron population, induced by weak activation of CsChrimson during two-photon imaging.



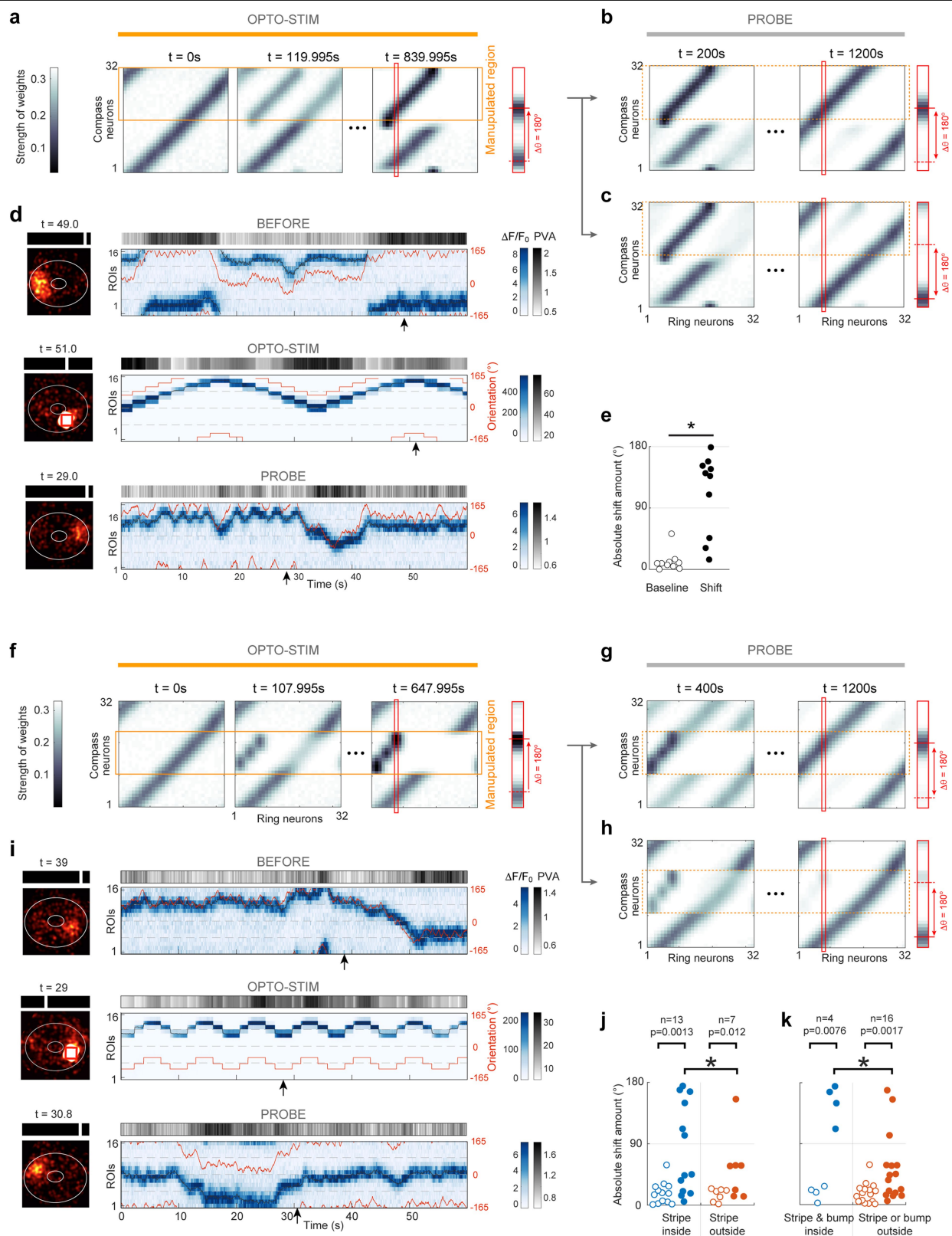
**Extended Data Fig. 2 | Simulation showing the mapping of a complex scene onto a stable heading representation and optogenetic bump offset shifting.**

**a**, A complex one-dimensional scene was generated via a mixture of four von Mises functions with random mean directions and random concentration parameters, shown for  $t = 0$ . **b, c**, Model simulation. Ring-neuron population activity (**b**, top) serves as the assumed source of visual input. A time series of angular velocity obtained from tethered flight data was used to compute movement of the visual scene. **b**, Bottom, compass-neuron population activity during simulated orientation. **c**, Time-varying synaptic weights between ring and compass neurons. The simulation began with random synaptic weights (left) and random initial activity of compass-neuron population. Ring attractor dynamics ensures a stable bump, albeit with a random offset. The initial turning of bump is not enforced by visual cues but by the angular velocity signal from tethered flight data. The same 400-s turning signal was repeated 3 times (Supplementary Information). Synaptic weights stabilize over time (**c**, right). After learning, a vertical cross-section of the stabilized synaptic weight matrix resembles the model ring-neuron activity profile shown in **a**. **d**, Simulation of optogenetic shift in offset. The simulation began with the stable mapping shown in **c**. **e**, During the probe trial, the newly mapped offset was consolidated. All simulation results shown are based on a post-synaptically gated plasticity rule, unless otherwise stated. Extended Data Figures 5, 6 and Supplementary Information provide the differences in predictions made by post- and pre-synaptically gated plasticity rules.



**Extended Data Fig. 3 | Bump dynamics after a closed-loop two-stripe manipulation.** **a–c**, Simulation of the time evolution of the synaptic weight matrix, induced by a visual scene with two vertical stripes. Conventions are the same as in Extended Data Fig. 2. **a**, The simulation began with the stabilized synaptic weight matrix shown in Fig. 2e. Visual input provided was two narrow von Mises functions, separated by  $180^\circ$ . Ring attractor dynamics ensured that the compass-neuron population maintained a single bump. Over time, the synaptic weight matrix develops two distinct bands of weak synapses (right), representing weakened connections from two active sets of ring neurons to a compass-neuron bump. **b, c**, When the system is then presented with a visual scene that has only one vertical stripe, there are two possible outcomes: ring attractor dynamics stabilizes an offset that is either shifted  $180^\circ$  from the original offset (**b**) or the same as the original offset (**c**). **d–i**, Natural bump-offset shifting with two identical vertical stripes (no optogenetic manipulation) separated by  $165^\circ$  in a  $330^\circ$  arena. **d–f**, Segments (60 s) of compass-neuron calcium transients before (**d**), during (**e**) and after (**f**) manipulation. Conventions are the same as in Fig. 2d, except that the red line represents the position of either one (**d, f**) or two (**e**) stripes. Imaging snapshots shown in the left panels were taken at times indicated with arrows beneath right panels. The bump offset is shifted by  $180^\circ$  in **f**, relative to its position in **e**

(Supplementary Video 4). **g**, Distribution of the absolute shift in offset measured across trials from all flies. Left, baseline variance; change in offset across two trials before manipulation. Right, baseline variance; change in offset across two trials after manipulation. Centre, change in offset across two trials separated by a manipulation trial. In three cases ( $n = 19$ ), the shift in offset was close to  $180^\circ$ . Unlike in simulations, in most two-stripe trials the bump position covers only half of the ellipsoid body because of the circular symmetry of the stimulus, which may underlie the apparently low yield of shifting (but see **h** and **i**; see Supplementary Information for further discussion). **h**, The number of bumps during the initial 15 s of 16 trials that did not exhibit a shift of  $180^\circ$  was significantly greater in trials that immediately followed a manipulation trial (red) than in a subsequent trial (blue) (bootstrap test of the mean difference, one-sided,  $P = 0.0004$ ), indicating that initial competition between two bumps eventually stabilizes to a single bump. This implies that the manipulation trial generated two competing offsets. **i**, The deviation of the bump offset during the initial 15 s relative to the average bump offset during final 30 s of the same trial was also significantly greater in the trial immediately following a manipulation trial than in a subsequent trial (bootstrap test of mean difference, one-sided,  $P = 0.0036$ ), which is a natural consequence of competition between two alternating bumps before one stabilizes.



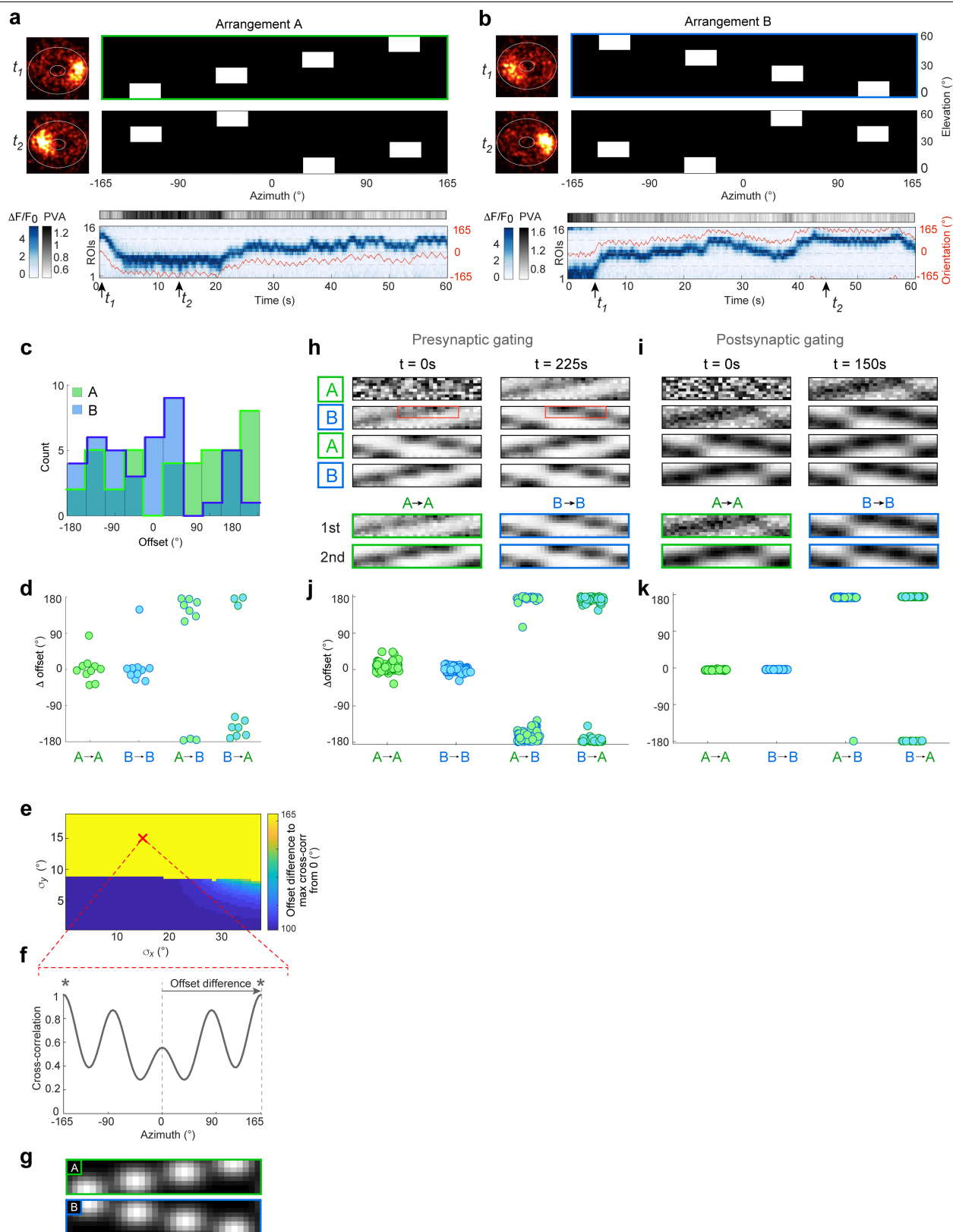
Extended Data Fig. 4 | See next page for caption.



**Extended Data Fig. 4 | Global offset shift by local optogenetic manipulation.**

The conventions are the same as in Extended Data Fig. 2. **a–e**, Local optogenetic manipulation spanning 180°. **a**, The simulation begins with a stabilized synaptic weight matrix, shown in Fig. 2e. Over time, a new map spanning 180° replaced approximately half of the original map (right). A portion of the synaptic weight matrix, corresponding to visual orientations that were not presented, was erased over time (top right corner of right panel). **b, c**, After manipulation, two potential maps (the original map and the newly imposed map) compete. Which map it is that eventually stabilizes and strengthens depends on whether or not the bump and stimulus begin in the newly mapped region of the ellipsoid body in the trial that immediately follows manipulation. **d**, Compass-neuron calcium transients before (top), during (middle) and after (bottom) optogenetic manipulation spanning 180° of the visual scene and the ellipsoid body. The conventions are the same as in Fig. 2d. Compare the offsets in the top and bottom panels. **e**, Distribution of the absolute shift in offset, measured across flies. White dots, baseline before manipulation; black dots, offset shift by manipulation (10 flies, bootstrapped mean difference test, one-sided,  $*P < 0.0001$ ). **f–k**, Local optogenetic manipulation spanning 60°. **f**, The simulation begins with the stabilized synaptic weight matrix shown in Fig. 2e. Over time, the newly imposed map replaces a portion of original map, which spans more than 60° because of the non-zero width (118° tail to tail) of the bump (bottom right). **g, h**, After the manipulation, two potential maps (the original map and the newly imposed map) compete. After the epoch of manipulation, if the bump begins in the manipulated region (**g**), the new map is likely to dominate and eventually strengthen. **i–k**, Optogenetic manipulation

spanning 60° of the visual scene and the ellipsoid body. **i**, Segments (60 s) of compass-neuron population activity before (top), during (middle) and after (bottom) manipulation. The position of stripe (bottom) is not in the manipulated domain, yet the bump is shifted to the optogenetically imposed offset (compare the offsets in the top and bottom panels). **j**, Left, data from 60°-span manipulation, after which a closed-loop probe trial begins with the stripe in the position that was sampled during manipulation. Open dots, baseline variance of the offset around mean, before manipulation. Solid blue dots, shift in offset induced by 60°-span manipulation. Across the population, the shift was significant (bootstrapped mean comparison, one-sided,  $P < 0.0013$ ). Right, data from 60°-span manipulation, after which closed-loop probe trial begins with the stripe outside the set of positions sampled during manipulation. Open dots, baseline variance. Solid red dots, shift in offset induced by manipulation. The shift was only marginally significant across the population (bootstrapped mean comparison, one-sided,  $P = 0.012$ ). The global extrapolation of local manipulation was facilitated when the stripe began in manipulated positions in the probe trial (binomial exact test,  $*P = 0.0059$ ) (Methods). **k**, Same data as in **j** but re-categorized. Left, in probe trials, both the bump and stripe began in a position sampled during the manipulation (4 out of 20 flies). All 4 flies showed a greater-than-90° shift during probe trials. Right, all other conditions (16 out of 20 flies). In total, 3 out of 16 flies showed a greater-than-90° shift. The facilitation of global extrapolation when both the bump and stripe began in manipulated positions was significant (binomial exact test,  $*P = 0.0012$ ) (Methods).

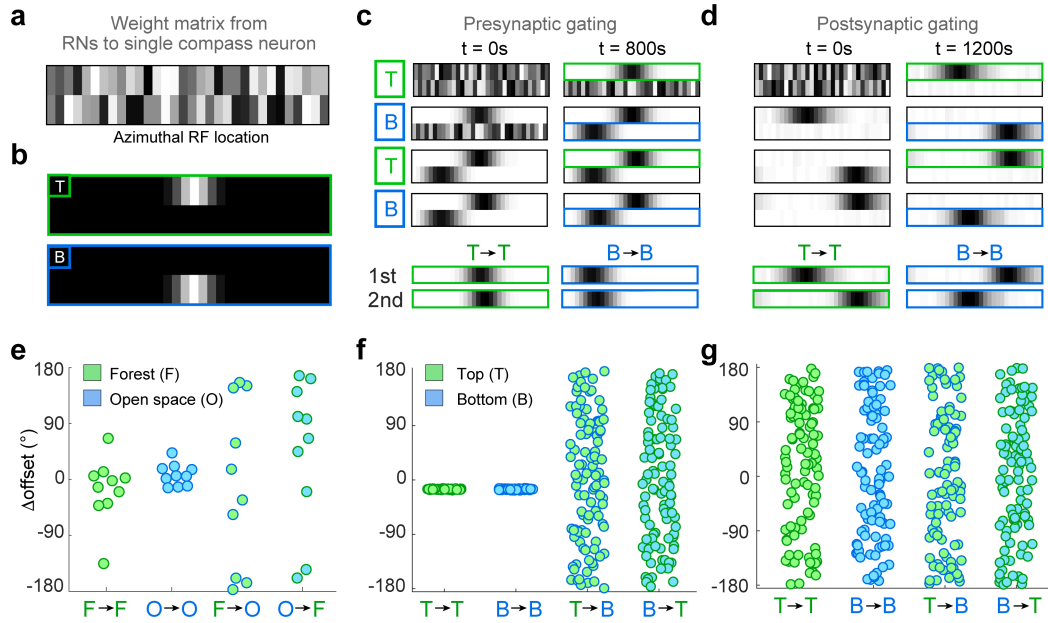


Extended Data Fig. 5 | See next page for caption.

**Extended Data Fig. 5 | Deterministic offset difference between two artificial scenes with the same local feature but different two-dimensional organization.**

The Supplementary Information provides a detailed discussion. **a**, Compass-neuron calcium transients measured during closed-loop tethered flight in an artificial scene, arrangement A (A). The conventions are the same as in Fig. 1h. **b**, Calcium transients from the same fly as in **a**, but with a different artificial scene, arrangement B (B). **c**, Distribution of the mean offset of each trial, pooled across all flies (Methods). Distributions of offsets relative to scenes A and B were not significantly different from uniform ( $n = 40$  trials from 10 flies, unimodality test by randomization,  $P = 0.0819$  for A,  $P = 0.1525$  for B). Compare with Fig. 1j. **d**, Distribution of offset shifts between two trials. The distribution of offset shifts between two artificial scenes, measured across flies, was significantly different from uniform distribution (unimodality test by randomization, from A to B,  $n = 10$  flies,  $P < 0.0001$ ; from B to A,  $n = 10$  flies,  $P < 0.0001$ ). The shift in offset was similar across different encounters with same scene, indicating that the offset was stable (unimodality test by randomization, from A to A,  $n = 10$  flies,  $P = 0.0001$ ; from B to B,  $n = 10$  flies,  $P = 0.0004$ ). Compare with Extended Data Fig. 6e. **e**, Parameter sweep to explore how two-dimensional Gaussian filters of different s.d., applied to the artificial scenes in **a** (arrangement A) and **b** (arrangement B), would affect shifts in offset between the two scenes. Filters represent the simplified effect of ring-neuron filtering of scenes. Shifts in offset should approximately match azimuthal shifts that would produce the best match (that is, maximum two-dimensional cross-correlation) between the filtered scenes. Each axis represents increasing s.d. of the applied two-dimensional Gaussian filter (**g**). The point marked with a red X is shown in **f**. **f**, Two-dimensional cross-correlation between two scenes in **a** and **b** after applying two-dimensional Gaussian filtering with  $15^\circ$  s.d. (red X in **e**). This filter size corresponds to a  $30^\circ$

full-width at half-maximum receptive field, which matches the average size of the minor axis of ellipses that fit ring-neuron receptive fields<sup>13,39</sup>. Higher filter sizes up to  $60^\circ$  full-width at half-maximum (the average size of the major axis of elliptical fits of ring-neuron receptive fields<sup>13,39</sup>) require similar azimuthal shifts to obtain a best match between the scenes (not shown in **e**). The azimuthal shift for the best match for this range of filters is  $165^\circ$ , a half rotation of the scene on the visual arena (as observed in **d**). **g**, Scenes in **a** and **b** after applying Gaussian filtering with  $15^\circ$  s.d. **h**, **i**, Simulation of pre- and post-synaptically gated plasticity rules applied when the model network is exposed to the two different filtered scenes shown in **g**. **h**, Evolution of the synaptic weight matrix with a pre-synaptically gated plasticity rule. Top left, initial random synaptic weight matrix from  $8 \times 32$  ring neurons to 1 of 32 compass neurons. Top right, after exposure to scene A. Each compass neuron responds most to a snapshot of the scene at a particular orientation. Second row, after exposure to scene B, a new snapshot is mapped to the compass-neuron heading representation. The locations of the top two horizontal bars in arrangements A and B overlap (red rectangles), which corresponds to a  $165^\circ$  shift in the two-dimensional cross-correlation in **e** and **f** (or a  $180^\circ$  shift in the  $360^\circ$  arena in simulations). This deterministic offset shift results in the same pinning offset and a retrieval of the same heading representation as before when the scene is repeated later (bottom two rows). The third and fourth rows show repeated exposure to scenes A and B. Bottom two rows, retrieval of the original offset. **i**, Evolution of the synaptic weight matrix with post-synaptically gated plasticity rule. The result is almost identical to **h**, given that all ring neurons and compass neurons are activated during simulation. **j**, **k**, Simulated offset shifts with pre-synaptically (**j**) and post-synaptically (**k**) gated plasticity rules. For each rule, 100 simulations were performed. Both the pre-synaptic and the post-synaptic rules reproduced the population data in **d**.



#### Extended Data Fig. 6 | Memory capacity of different plasticity rules.

**a–d**, Simulation of pre- and post-synaptically gated plasticity rules with simple two-dimensional scenes. **a**, Initial random synaptic weight matrix from  $2 \times 32$  ring neurons to 1 of 32 compass neurons. **b**, Two simple simulated scenes activate mutually exclusive ring neurons. T, top ring neurons are active; B, bottom ring neurons are active. **c**, Evolution of synaptic weights for a pre-synaptically gated plasticity rule. Top left, initial random weight matrix before presenting scene T. Top right, after exposure to scene T, only synapses from active ring neurons (top row of ring neurons in **e**) were updated, while synapses from all other ring neurons (bottom row of ring neurons in **e**) remained intact. Second row, after exposure to scene B, ring neurons that were previously inactive became activated, and their synapses were updated. Third row, when scene T was presented again, the offset between scene orientation and bump position was the same as when scene T was first presented (**f**). **d**, Evolution of

synaptic weights for a post-synaptically gated plasticity rule. Synapses from inactive ring neurons are erased upon each encounter with a new scene. This would shift offset across two encounters of the same scene if the fly experiences a different scene between them. **e**, Population data are from ten flies. Distribution of offset shifts between two trials in Fig. 1h, i. The distribution of offset shifts between two different natural scenes, measured across flies, is not significantly different from uniform distribution (unimodality test by randomization, from F to O,  $P = 0.489$ ; from O to F,  $P = 0.1504$ ). Different encounters of the same scene lead to similar, near-zero offset shifts, indicating stability of offset (unimodality test by randomization, from F to F,  $P = 0.0035$ ; from O to O,  $P < 0.0001$ ). **f, g**, Simulated offset shifts with pre-synaptically (**f**) and post-synaptically (**g**) gated plasticity rules. For each rule, 100 simulations were performed.

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☒ ☐ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection

Image collection (Scanimage 2016b), Behavioral image collection (StreamPix 6), Behavioral data collection (Matlab 2018b with NIDaq)

Data analysis

Data analyses (Matlab R2018b)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data and code are freely available without restriction at [http://research.janelia.org/jayaraman/Kim\\_etal\\_Nature2019\\_Downloads/](http://research.janelia.org/jayaraman/Kim_etal_Nature2019_Downloads/)

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- ☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)



# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The number of flies to collect was chosen to be 10, which is a sufficient number to perform non-parametric testings, for all experimental and control groups, unless mentioned otherwise in Methods.
Data exclusions	No exclusion of flies. In each fly, segments when fly was not flying were excluded from analyses.
Replication	The optogenetic bump offset shifting has been replicated ten times with different conditions, all of which reproduced significant effects. Five of them are reported in the manuscript.
Randomization	All flies in this study were randomly selected from their housing vials.
Blinding	Not applicable.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	5- to 8-day old female <i>Drosophila melanogaster</i>
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve samples collected from the field.
Ethics oversight	No ethical approval or guidance was required for <i>Drosophila</i> physiological experiments.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Gene expression cartography

<https://doi.org/10.1038/s41586-019-1773-3>

Received: 1 February 2019

Accepted: 7 October 2019

Published online: 20 November 2019

Mor Nitzan<sup>1,2,3,6</sup>, Nikos Karaïskos<sup>4,6</sup>, Nir Friedman<sup>3,5\*</sup> & Nikolaus Rajewsky<sup>4\*</sup>

Multiplexed RNA sequencing in individual cells is transforming basic and clinical life sciences<sup>1–4</sup>. Often, however, tissues must first be dissociated, and crucial information about spatial relationships and communication between cells is thus lost. Existing approaches to reconstruct tissues assign spatial positions to each cell, independently of other cells, by using spatial patterns of expression of marker genes<sup>5,6</sup>—which often do not exist. Here we reconstruct spatial positions with little or no prior knowledge, by searching for spatial arrangements of sequenced cells in which nearby cells have transcriptional profiles that are often (but not always) more similar than cells that are farther apart. We formulate this task as a generalized optimal-transport problem for probabilistic embedding and derive an efficient iterative algorithm to solve it. We reconstruct the spatial expression of genes in mammalian liver and intestinal epithelium, fly and zebrafish embryos, sections from the mammalian cerebellum and whole kidney, and use the reconstructed tissues to identify genes that are spatially informative. Thus, we identify an organization principle for the spatial expression of genes in animal tissues, which can be exploited to infer meaningful probabilities of spatial position for individual cells. Our framework ('novoSpaRc') can incorporate prior spatial information and is compatible with any single-cell technology. Additional principles that underlie the cartography of gene expression can be tested using our approach.

Single-cell RNA sequencing (scRNA-seq) has revolutionized our understanding of the rich heterogeneous cellular populations that make up tissues, the dynamics of developmental processes and the underlying regulatory mechanisms that control cellular function<sup>1–4</sup>. However, to understand how single cells orchestrate multicellular functions, it is crucial to have access not only to the identities of single cells but also to their spatial context. This is a challenging task, as tissues must commonly be dissociated into single cells before scRNA-seq can be performed, and thus the original spatial context and relationships between cells are lost. Two seminal papers tackled this problem computationally<sup>5,6</sup>—the key idea being to use a reference atlas of informative marker genes as a guide to assign spatial coordinates to sequenced cells. This concept was successfully used in various tissues<sup>7–11</sup>, including the early *Drosophila* embryo<sup>12</sup>. However, such methodologies rely heavily on the existence of an extensive reference database for spatial expression patterns, which may not always be available or straightforward to construct. Moreover, in practice the number of available reference marker genes is usually not large enough to label each spatial position with a distinct combination of reference genes, making it impossible to uniquely resolve cellular positions. More generally, marker genes, even when available, convey limited information, which could possibly be enriched by the structure of single-cell data.

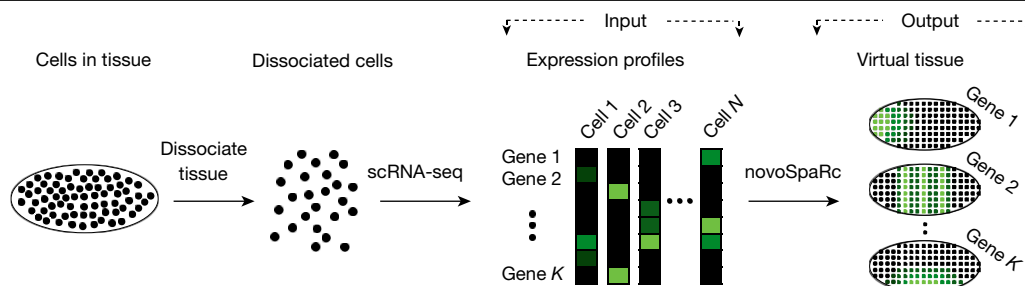
To this aim, we developed a new computational framework (novoSpaRc), which allows for de novo spatial reconstruction of single-cell gene expression, with no inherent reliance on any prior information, and the flexibility to introduce it when it does exist (Fig. 1). Similar to solving a puzzle, we seek the optimal configuration of pieces (cells)

that recreates the original image (tissue). However, contrary to a typical puzzle, here we do not have access to the image that we aim to reconstruct. Although the number of ways to spatially arrange (or 'map') sequenced cells in tissue space is enormous, our hypothesis is that gene expression in the vast majority of these arrangements will not be as organized as in the real tissue. For example, we know that typically there exist genes that are specifically expressed in spatially contiguous territories and are thus consistent with only a small subset of all possible arrangements. We therefore set out to identify simple, testable assumptions that govern how gene expression is organized in space, and to subsequently find the arrangements of cells that best respect those assumptions.

## novoSpaRc charts gene expression in tissues

Here, we specifically explore the assumption that cells that are physically close tend to share similar transcription profiles, and vice versa (Extended Data Fig. 1a, Supplementary Methods). Biologically, this phenotype can result from multiple mechanisms, such as gradients of oxygen, morphogens and nutrients, the trajectory of cell development and communication between neighbouring cells. We stress that this is an assumption about overall gene expression across the entire tissue—not about individual genes and not about all cells that are physically close (Supplementary Methods). We show that, on average, the distance between cells in expression space increases with their physical distance, for diverse tissues in mature organisms or whole embryos in early development. Thus, to predict the spatial locations of sequenced cells, we

<sup>1</sup>John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. <sup>2</sup>Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>3</sup>School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel. <sup>4</sup>Systems Biology of Gene Regulatory Elements, Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany. <sup>5</sup>Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel. <sup>6</sup>These authors contributed equally: Mor Nitzan, Nikos Karaïskos. \*e-mail: nir.friedman@mail.huji.ac.il; rajewsky@mdc-berlin.de



**Fig. 1 | Overview of novoSpaRc.** A matrix that contains single-cell transcriptome profiles, sequenced from dissociated cells, is the main input for novoSpaRc. The output is a virtual tissue of a chosen shape, which can be queried for the expression of all genes quantified in the data.

seek to find a map of sequenced cells to tissue space ('cartography') such that overall structural correspondence is preserved—meaning that, overall, cells have similar relative distances to other cells in expression and physical space. The physical space is anchored by locations that may be either known (such as the reproducible cellular locations in the early stages of development of the *Drosophila* embryo<sup>13</sup>) or approximated by a grid (Supplementary Methods). The distances are first computed for each pair of cells across graphs constructed over the two spaces, to account for the underlying structure of the data (Supplementary Methods). Then, novoSpaRc optimally aligns the distances of pairs of cells between the expression data and geometric features of the physical space, in a way that is consistent with spatial expression profiles of marker genes when these are available (Methods, Supplementary Methods). For reasons that are both biologically and computationally motivated, we seek a probabilistic mapping that assigns each cell a distribution over locations on the physical space (Supplementary Methods). We formulate this as a generalized optimal-transport problem<sup>14–16</sup>, which has been proven to be increasingly valuable for diverse fields (including biology<sup>17,18</sup>) and renders the task of reconstruction feasible for large datasets. Specifically, we formulate an interpolation between entropically regularized Gromov–Wasserstein<sup>19,20</sup> and optimal-transport<sup>21</sup> objectives, which serves to satisfy the assumption of structural correspondence between gene expression space and physical space, and to match prior knowledge when available (Methods). We show that this optimization problem can be efficiently solved using projected gradient descent reduced to iterations of linear optimal-transport sub-problems (Supplementary Methods). To systematically assess the performance of novoSpaRc, we used a simple generative model of spatial gene expression to show that it can robustly recover it (Supplementary Methods, Extended Data Fig. 1b–d).

### novoSpaRc reconstructs tissues de novo

Focusing on real single-cell datasets, we first reconstructed tissues de novo that have inherent symmetries that render them effectively one-dimensional, such as the mammalian intestinal epithelium<sup>10</sup> and liver lobules<sup>7</sup>. Schematic figures of the reconstruction process are shown in Fig. 2a, e. Cells were previously classified into seven distinct zones for the intestine, or nine layers for the liver, on the basis of robust marker gene information<sup>7,10</sup>. We found that the average pairwise distances between cells in expression space increased monotonically with the pairwise distances in physical one-dimensional space (Fig. 2b, f), consistent with our structural correspondence assumption.

We used novoSpaRc to embed the expression data into one dimension. The embedded coordinates of single cells correlated well on average with their layer or zone memberships (Fig. 2c, g, Supplementary Methods). The median Pearson correlation coefficient for reconstructed expression patterns to original patterns for the top 100 variable genes was 0.99 for intestine and 0.94 for liver (Supplementary Methods), and the fraction of cells that were correctly assigned up to one layer away from their original layer was 0.98 for intestine and 0.73

for liver (Supplementary Methods, Extended Data Fig. 2a, b). novoSpaRc captured spatial expression patterns of the top zonated genes and spatial division of labour within the intestinal epithelium—as well as within the layers of the liver lobules (Methods, Fig. 2d, h, Extended Data Fig. 3a, b), in which cells in different tissue layers perform different tasks and exhibit different expression profiles. For the intestine, varying the grid resolution to include either fewer or more embedded zones did not compromise the quality of the reconstructed expression patterns (Extended Data Fig. 3c), which shows the potential for increased resolution of single-cell-based relative to atlas-based embedding.

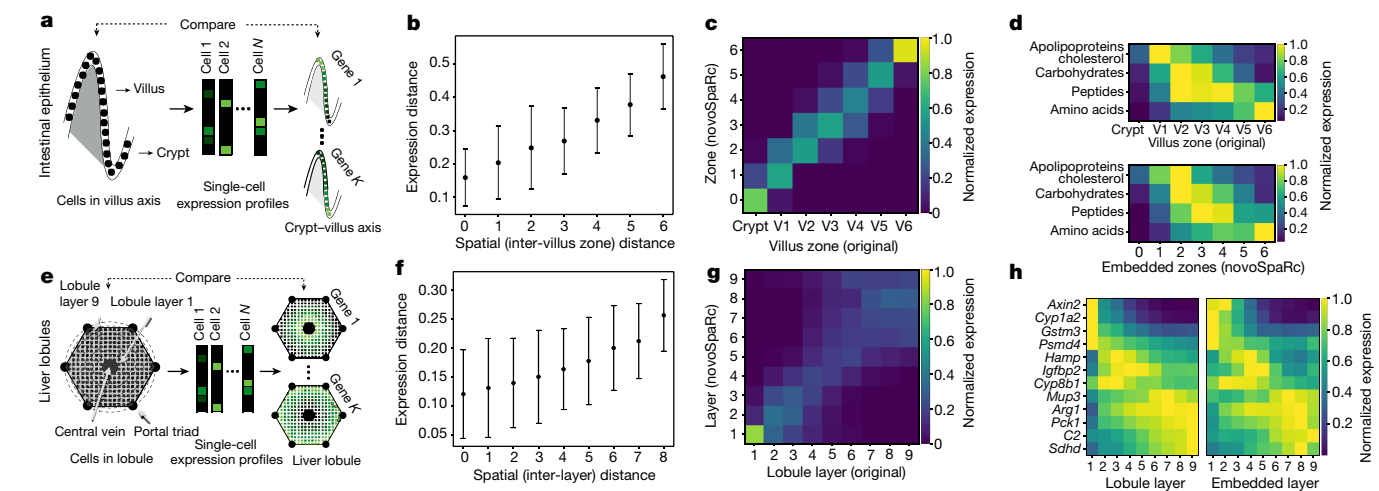
### novoSpaRc reconstructs early embryos

Next, we focused on spatially reconstructing the well-studied *Drosophila* embryo, as a more-challenging, higher-dimensional tissue. Late in stage 5 of development, the fly embryo consists of around 6,000 cells. It has been previously suggested<sup>22</sup> that at early stages of fly development, the expression levels of gap genes can be optimally decoded into positional information. The expression levels of 84 transcription factors were quantitatively registered using fluorescence in situ hybridization (FISH) for each of the cells by the Berkeley *Drosophila* Transcription Network Project (BDTNP)<sup>13</sup>.

To assess the performance of novoSpaRc, we first simulated scRNA-seq data by in-silico dissociating the BDTNP dataset into single cells (Methods), and then attempted to reconstruct the original expression patterns across the tissue both de novo and by using marker genes (Fig. 3a). Similarly to the 'one-dimensional' datasets, we found a monotonically increasing relationship between the cell–cell pairwise distances in expression space and in physical space (Fig. 3b), confirming that the data adheres to our structural correspondence assumption.

The reconstructed patterns of spatial gene expression highly correlated with the original ones (Fig. 3c). We found that the novoSpaRc reconstruction that incorporated both structural and marker gene information outperformed the reconstruction based on only the latter, and that performance was saturated at two marker genes (Fig. 3c), independently of the marker genes used. As expected, the quality of the reconstruction increased with the number of genes used to provide structural information in expression space, and with the fraction of spatially informative genes (Supplementary Methods, Extended Data Fig. 4a, b). The majority of spatial patterns were recapitulated faithfully even when only a single marker gene was used (Fig. 3c, d). In addition, novoSpaRc identified the physical neighbourhoods from which cells originated when used de novo (up to inherent symmetries; see Supplementary Methods), and pinpointed their true locations ( $P < 0.05$  compared to random assignment) when a handful of marker genes were used (Fig. 3e, Extended Data Fig. 5a, b).

We examined the expression patterns of four transcription factors that span the dorsal–ventral and anterior–posterior axes (Fig. 3d). The quality of the reconstruction improved when applying the structural correspondence assumption (Supplementary Methods, Extended Data Fig. 5d). The de novo reconstruction correctly identified both axes of the embryo, and

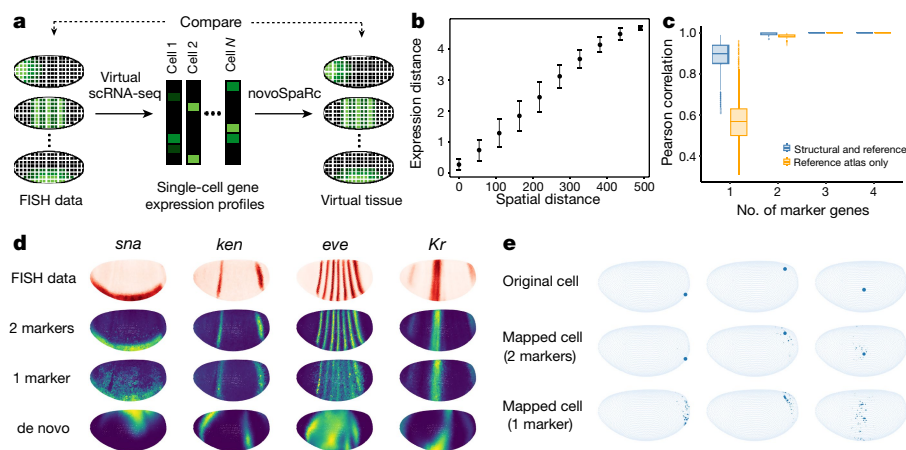


**Fig. 2 | novoSpaRc successfully reconstructs complex tissues with effective one-dimensional structure de novo.** **a, e,** The reconstruction scheme for the mammalian intestinal epithelium (**a**) and liver lobules (**e**). **b, f,** Demonstration of the monotonic relationship between cellular pairwise distances in expression and physical space for intestinal epithelium (**b**) and liver lobules (**f**). Distances are measured as weighted shortest paths along the graphs constructed over physical or expression spaces. Data are mean  $\pm$  s.d. **c, g,** novoSpaRc infers the original spatial context of single cells of the intestinal

epithelium (**c**) and liver lobules (**g**) with high accuracy. Heat maps show the inferred distribution over embedded layers (rows) for the cells in each of the original layers (columns). **d, h,** novoSpaRc captures the spatial division of labour of averaged expression of genes that have a role in the absorption of different classes of nutrients in the intestine (**d**) and the spatial expression patterns of a group of pericentral, periportal and non-monotonic genes in the liver lobule (**h**). The expression level of each gene in **d** and **h** is normalized to its maximum value.

the reconstructed portrait was remarkably similar to the original one (Fig. 3d). In general—because de novo reconstruction is performed without any prior information that would anchor the cells—the reconstructed configuration is similar up to global transformations (reflections, rotations and translations), relative to the respective axes of symmetry (Supplementary Methods). Consequently, the resulting patterns of gene expression might be shifted or flipped relative to the expected ones. However, there are features of a faithful reconstruction that we can test for, such that the reconstruction would be robust to small changes in the optimization parameters (Supplementary Methods, Extended Data Fig. 4i) and that

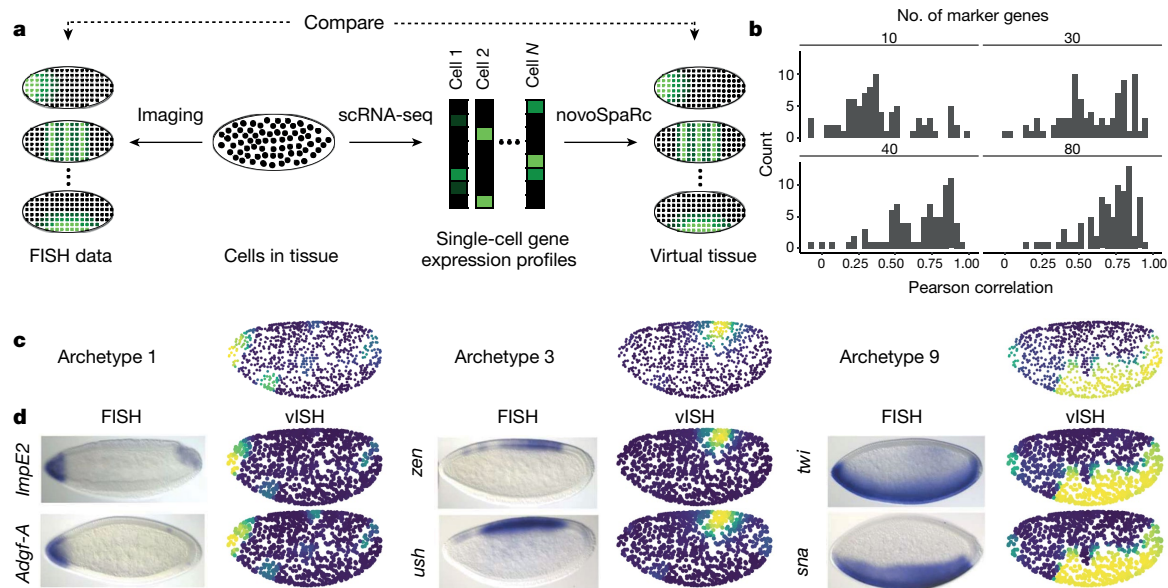
the embedding of cells onto the embryo would be relatively localized—as we would expect for a biologically meaningful embedding (Fig. 3e). This means that the distribution over locations that each cell is assigned should be localized, and indeed, the mean standard deviation of that distribution for all cells is significantly lower than that of a randomized embedding (Supplementary Methods, Extended Data Fig. 4j). Furthermore, we demonstrated that the results from novoSpaRc—as measured by correlation to observed imaging data and optimization error—were robust to optimization parameters and sources of noise, including partial sampling of cells, additive expression noise and dropouts (Extended Data Fig. 4c–h).



**Fig. 3 | novoSpaRc accurately reconstructs the *Drosophila* embryo on the basis of the BDTNP dataset<sup>13</sup>.** **a,** FISH data are used to create virtual scRNA-seq data, which novoSpaRc inputs to reconstruct a virtual embryo. **b,** Demonstration of the structural correspondence hypothesis. Pairwise cellular distances in expression space increase monotonically with distances in physical space. Data are mean  $\pm$  s.d. **c,** novoSpaRc spatially reconstructs the *Drosophila* embryo with only one marker gene. The quality of reconstruction (measured by Pearson correlation with FISH data) increases with the number of marker genes and saturates at perfect reconstruction at two marker genes, when using both structural information and marker gene information (blue boxes). This outperforms reconstruction that relies only on marker gene

information (yellow boxes). The results are averaged for 100 different combinations of marker genes. For the box plots, the centre line is the median, box limits are the 0.25 and 0.75 quantiles and whiskers extend to  $\pm 2.698$  s.d. **d,** Visualization of the reconstruction results for four transcription factors. The original FISH data (first row) are compared to reconstruction by novoSpaRc that exploits both structural and marker gene information (using two marker genes and one marker gene) and reconstruction without any marker gene information (de novo). **e,** The original locations of three cells are compared to their respective reconstructed locations by novoSpaRc (using two marker genes and one marker gene). The expression patterns of the marker genes used for the results in **d** and **e** are shown in Extended Data Fig. 5c.





**Fig. 4 | novoSpaRc identifies spatial archetypes in the *Drosophila* embryo by using scRNA-seq data.** **a**, Schematic overview. The expression patterns as reconstructed by novoSpaRc are compared with the BDTNP expression values. **b**, Reconstruction of the *Drosophila* embryo using scRNA-seq data. Distributions of gene-specific Pearson correlation coefficients reflect better

reconstruction with increasing number of marker genes. **c**, Three of the spatial archetypes (1, 3 and 9) that novoSpaRc identified in the *Drosophila* embryo. **d**, Representative genes for each of the spatial archetypes depicted in **c**. FISH data (left columns) are compared against the corresponding novoSpaRc predictions ('virtual in situ hybridization' (vISH); right columns).

We next used novoSpaRc to reconstruct the stage 6 *Drosophila* embryo by using a scRNA-seq dataset<sup>12</sup> (Fig. 4a). In that study, 84 marker genes were required to reconstruct a virtual embryo by distributing 1,297 cells over 3,039 locations. When we used novoSpaRc with the combination of both structural information and the reference atlas, the accuracy of reconstruction increased with the number of marker genes, reaching high correlation (Pearson correlation coefficient, 0.74) with the FISH data (Fig. 4b, Extended Data Fig. 5e). The de novo, atlas-free reconstruction accurately separated the major post-gastrulation spatial domains (mesoderm, neurogenic ectoderm and dorsal ectoderm), as well as finer spatial domains (Fig. 4c, d). We clustered the reconstructed patterns of the highly variable genes and averaged to obtain a representative pattern for each cluster, which we term the 'archetype' (Methods, Supplementary Information). novoSpaRc identified numerous distinct spatial archetypes (Fig. 4c, d, Extended Data Fig. 6). We compared representative genes of each spatial archetype with FISH images to visually assess the accuracy of the spatial reconstruction. Gene patterns that were expressed through the anterior–posterior or the dorsal–ventral axis were largely recapitulated: typical genes of the mesoderm (dorsal ectoderm), such as *twi* and *sna* (*zen* and *ush*), were colocalized ventrally (dorsally) (Fig. 4c, d, right, middle). novoSpaRc accurately captured localized spatial populations (Fig. 4c, d, left, Extended Data Fig. 6, archetype 5), whereas less-extensive spatial domains were reconstructed with varying degrees of accuracy (Extended Data Fig. 6). Note that within the de novo reconstruction, accurate localization entails global transformations, as described above (Supplementary Methods).

Before proceeding to more complex tissues, we reconstructed the zebrafish embryo dataset<sup>5</sup> (Extended Data Fig. 7). Similar to the original seminal study, we mapped the cells onto the surface of a hemisphere consisting of 64 distinct locations. The resulting spatial expression patterns highly correlated to the experimentally verified ones; novoSpaRc reconstructed the zebrafish embryo by using only 15 marker genes (in contrast to the 47 genes that were previously required<sup>5</sup>) and the accuracy of the reconstruction increased with the number of marker genes (Extended Data Fig. 7, Methods). Furthermore—in contrast to previous reconstructions—no data imputation or other specialized preprocessing was necessary<sup>5</sup>.

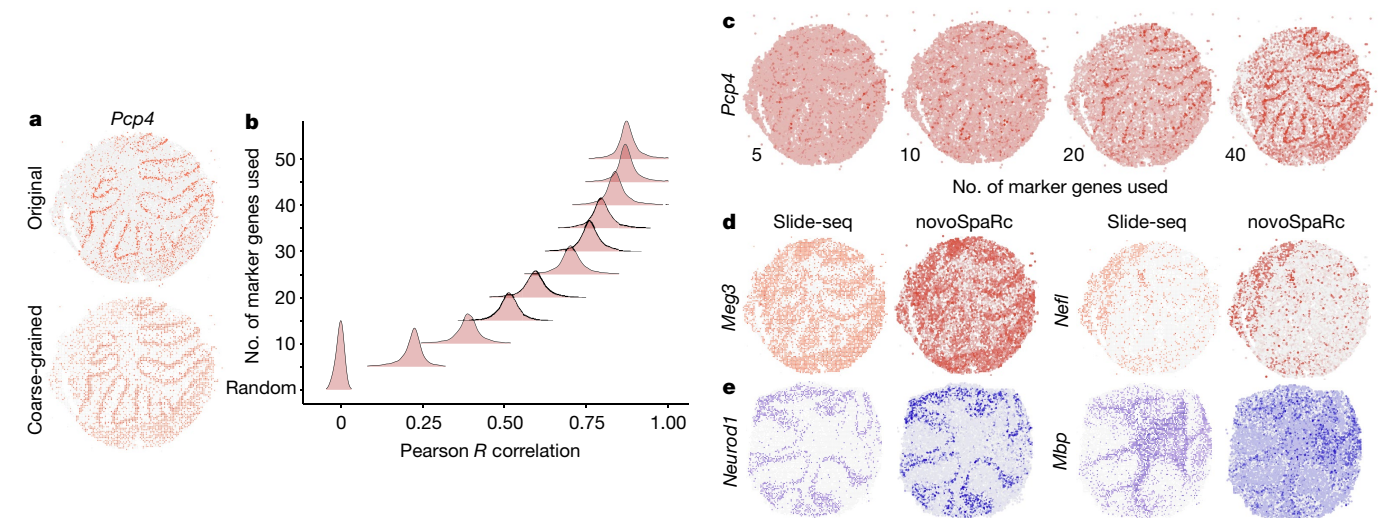
### novoSpaRc charts diverse complex tissues

To further demonstrate the applicability of novoSpaRc to complex tissues, diverse sequencing technologies and different organisms, we used it to reconstruct slices of mammalian brain cerebellum<sup>23</sup> (Fig. 5), the mammalian kidney<sup>24</sup> (Extended Data Fig. 8) and a dataset of hundreds of individual *Drosophila* embryos<sup>22</sup> (Extended Data Fig. 9).

The adult mammalian brain is a well-studied, highly differentiated and complex tissue. To benchmark the capabilities of novoSpaRc in reconstructing complex tissues, we used mouse cerebellum slices from a recently developed spatial transcriptomics technology<sup>23</sup>. The dataset of sagittal sections contained 46,376 locations, corresponding to a single cell or a few cells, with a median of 52 quantified transcripts per location. To provide enough information to novoSpaRc, we first coarse-grained the data by binning neighbouring locations. This resulted in 7,704 locations, with a median of 379 quantified transcripts per location (Methods, Fig. 5a). novoSpaRc successfully reconstructed the whole transcriptome, with a Pearson correlation coefficient of 0.5 over all 15,878 genes when using 15 marker genes and 0.94 when using 50 marker genes (Fig. 5b, Supplementary Methods). Spatial expression patterns emerged when using only a few markers. For example, spatial positions of Purkinje cells were revealed by reconstructing with only five marker genes (excluding all genes exhibiting an absolute Pearson correlation coefficient with *Pcp4* of 0.25 or higher). The signal improved markedly when more markers were included (Fig. 5c). The reconstructed cerebellum slices showed concordance with the original spatial gene expression for a large number of known cell-type marker genes (Fig. 5d). To illustrate the versatility of novoSpaRc, we further applied it to a coronal section of a brain cerebellum<sup>23</sup>, with similar results (Fig. 5e).

Next, we used novoSpaRc to spatially reconstruct a single-cell dataset from whole kidney<sup>24</sup>, which is a complex tissue with stereotypical organization. In the absence of a reference atlas of gene expression, the reconstruction was performed de novo. We focused on six major cell types of the kidney (Extended Data Fig. 8) and mapped the cells onto a two-dimensional target space. The de novo reconstruction recapitulated the urine flow within the kidney sub-compartments, as shown by the spatial gene expression of corresponding marker genes (Extended Data Fig. 8). We note that, as no prior information was required for this





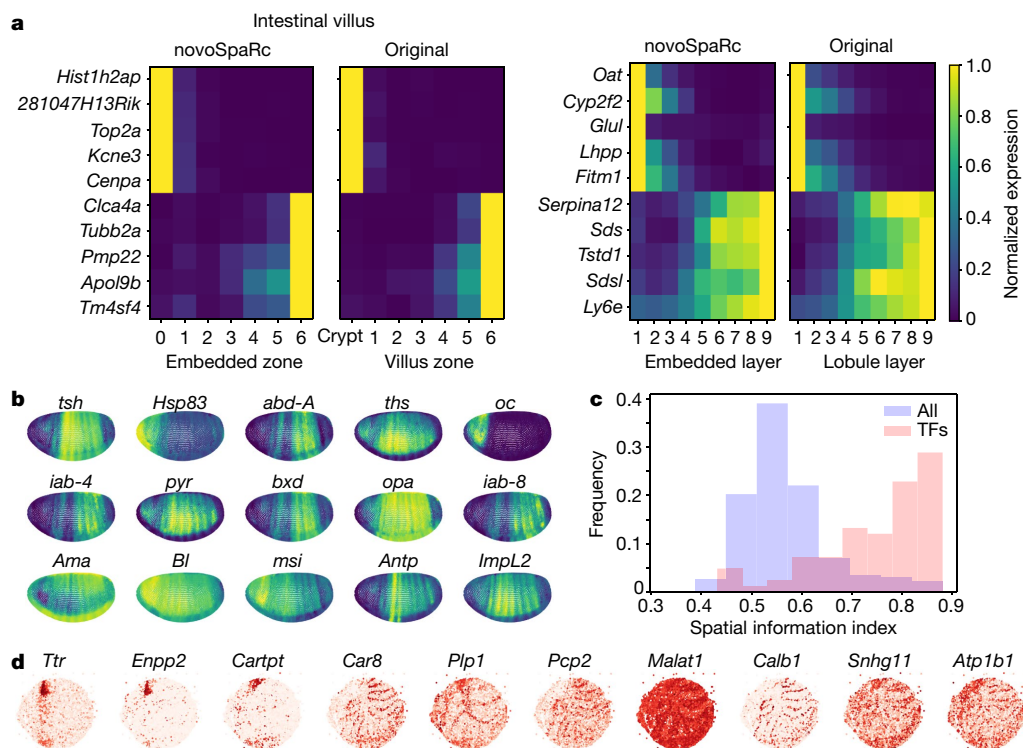
**Fig. 5 | novoSpaRc reconstructs mouse cerebellum tissue.** **a**, The original and the coarse-grained spatial expression of a marker of Purkinje cells (*Pcp4*) in a sagittal section of the cerebellum from direct spatial RNA sequencing<sup>23</sup>. **b**, The overall Pearson correlation between original gene expression and gene expression predicted by novoSpaRc increases markedly as more marker genes are used. The correlation when using only five marker genes is substantially

higher than that of a random mapping of cells to locations. Density plots contain values for all 15,878 genes. **c**, The spatial gene expression of *Pcp4* is visible with only five marker genes and is enhanced as more markers are used for the reconstruction. **d**, Examples of original and predicted expression for neuronal marker genes. Reconstruction was performed with 35 marker genes. **e**, novoSpaRc accurately reconstructs a coronal section of the cerebellum<sup>23</sup>.

reconstruction, this case demonstrates the applicability of novoSpaRc to a wide variety of medically relevant tissues.

Finally, to show that novoSpaRc can reconstruct not only a prototypic tissue but also individual samples, we used a dataset that captures

expression patterns in hundreds of individual *Drosophila* embryos<sup>22</sup>. In this dataset, the expression of four gap genes and four pair-rule genes was measured along the anterior–posterior axis for 101 and 177 embryos, respectively, providing a distribution over expression



**Fig. 6 | novoSpaRc identifies spatially informative genes.** **a**, Identifying spatially informative genes in the mammalian intestine and liver. We identify de novo (that is, with no marker genes used) the most highly zoned genes along the crypt-to-villus axis in the intestine (left) and across the axis of a liver lobule (right). The prediction of novoSpaRc is compared against the original expression patterns. The expression level of each gene is normalized to its maximum value. **b–d**, Identifying spatially informative genes in the *Drosophila* embryo (reconstruction with the BDTNP marker genes) and a slice of the

mammalian cerebellum (reconstruction with 50 markers), using a measure of spatial autocorrelation. **b**, Expression patterns of the top 15 spatially informative genes in the *Drosophila* embryo. **c**, The spatial autocorrelation values (spatial information index) of the 84 transcription factors (TFs) chosen for the BDTNP dataset<sup>13</sup> are among the highest values over all 8,924 genes of the fly embryo, demonstrating that they are identified to be highly spatially informative. **d**, Top 10 spatially informative genes (out of the top 1,000 variable genes) in a section of the cerebellum.

patterns. We used novoSpaRc to reconstruct the expression patterns of the gap and pair-rule genes for individual embryos. For a given embryo, novoSpaRc reconstruction using a reference atlas based on the gene expression within the same embryo consistently outperformed reconstruction using a reference atlas based on the averaged gene expression across all embryos in the dataset (Extended Data Fig. 9)—yet reached high correlation values for both (median Pearson correlation coefficients for reconstructing a fourth gene based on the three remaining genes were 0.99 (for expression within the same embryo) (0.95 for expression averaged across embryos) and 0.94 (0.77) for the gap and pair-rule genes, respectively).

We examined the effect of the interpolation between structural and marker gene information, and evaluated the performance of novoSpaRc by comparing it to available reconstruction methods that fully rely on a reference atlas (Seurat<sup>5</sup> and DistMap<sup>12</sup>) (Extended Data Figs. 10, 11). novoSpaRc has several advantages when compared to the other existing methods and overall shows substantial benefits in reconstruction performance (Extended Data Fig. 10, Supplementary Discussion).

## Identifying spatially informative genes

A novoSpaRc-based spatial reconstruction allows us to identify known and potentially new spatially informative genes directly from the single-cell sequencing data. For the intestine and liver datasets, we recovered highly zonated genes without a reference atlas (Methods, Supplementary Information), and found that the top inferred zonated genes were supported experimentally and/or computationally (Fig. 6a, Supplementary Tables 1, 2). Gene ontology enrichment analysis<sup>25</sup> further revealed that zonation-compatible biological processes enriched for different domains in the intestine and the liver were reconstructed by novoSpaRc (Supplementary Information). For the *Drosophila* single-cell dataset, we ranked all 8,924 genes according to their spatially informative rank (Methods, Fig. 6b, Supplementary Information), and found that transcription factors were (as known from classic genetics<sup>26</sup>) among the most highly informative genes (Fig. 6c). In addition, novoSpaRc identified numerous long non-coding RNAs and transcription factors as being highly spatially informative, many of them already predicted in a previous study<sup>12</sup>. Finally, we ranked all 15,878 genes in the cerebellum by their spatially informative rank (Methods, Fig. 6d, Supplementary Information), and found that well-known marker genes with a defined pattern of spatial expression are indeed among the highest-ranking spatially informative genes (Fig. 6d).

## Discussion

Together, we have demonstrated here that one can spatially reconstruct diverse biological tissues on the basis of a simple hypothesis about how gene expression is organized in space—a structural correspondence between the distances of cells in expression space and in physical space—and that it can be used to extract spatially informative genes. Our current implementation is based on pairwise comparison of cells and locations. This requirement can be readily altered. In fact, it is compelling to hypothesize that within certain biological contexts, different cell types may require higher-order interactions or exhibit different principles of spatial organization. Furthermore, we stress that because of the availability of general mathematical results in optimal-transport theory, our framework is versatile and can support a variety of alternative ways to compare distances in expression and physical space by varying the optimization loss functions (Methods, Supplementary Methods). Such alternative schemes are not currently supported by novoSpaRc, but could be implemented.

Our data analyses and the success of the reconstructions by novoSpaRc suggest that we have identified a general principle for how gene expression is organized in tissue space (Supplementary Discussion). It will be interesting to find tissues for which this organization principle

is weak or not valid. However, this principle may be underestimated, as most of the single-cell data available are relatively shallow and noisy. Our data also suggest that many more genes than perhaps anticipated are involved in spatial features and functions (including physiology and pathophysiology) of tissue. We have demonstrated that we can systematically identify at least a subset of these genes directly from single-cell data. In the future, we will extend these analyses to identify genes that are predicted to functionally interact in space. Finally, our developed framework can be flexibly extended beyond spatial reconstruction. We are currently using it to recover different types of biological signals, such as temporal progression on short (for example, cell cycle) and long (for example, developmental) timescales.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1773-3>.

- Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14**, 618–630 (2013).
- Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34**, 1145–1160 (2016).
- Altschuler, S. J. & Wu, L. F. Cellular heterogeneity: do differences make a difference? *Cell* **141**, 559–563 (2010).
- Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Mol. Cell* **58**, 610–620 (2015).
- Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
- Achim, K. et al. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.* **33**, 503–509 (2015).
- Halpern, K. B. et al. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature* **542**, 352–356 (2017).
- Durruthy-Durruthy, R. et al. Reconstruction of the mouse otocyst and early neuroblast lineage at single-cell resolution. *Cell* **157**, 964–978 (2014).
- Waldhaus, J., Durruthy-Durruthy, R. & Heller, S. Quantitative high-resolution cellular map of the organ of Corti. *Cell Rep.* **11**, 1385–1399 (2015).
- Moor, A. E., et al., Spatial reconstruction of single enterocytes uncovers broad zonation along the intestinal villus axis. *Cell* **175**, 1156–1167 (2018).
- Habib, N. et al. Div-Seq: Single-nucleus RNA-seq reveals dynamics of rare adult newborn neurons. *Science* **353**, 925–928 (2016).
- Karaiskos, N. et al. The *Drosophila* embryo at single-cell transcriptome resolution. *Science* **358**, 194–199 (2017).
- Berkeley *Drosophila* Transcription Network Project. <http://bdnptn.lbl.gov:8080/Fly-Net/>.
- Monge, G. Mémoire sur la théorie des déblais et des remblais. *Historie de l'Académie Royale des Sciences de Paris* **1781**, 666–704 (1781).
- Villani, C. *Topics in Optimal Transportation* (American Mathematical Society, 2003).
- Villani, C. *Optimal Transport: Old and New* Vol. **338** (Springer, 2008).
- Schiebinger, G. et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* **176**, 928–943 (2019).
- Forrow, A. et al. Statistical optimal transport via geodesic hubs. Preprint at <https://arxiv.org/abs/1806.07348> (2018). If ref. 18 (preprint) has now been published in final peer-reviewed form, please update the reference details if appropriate.
- Mémoli, F., On the use of Gromov–Hausdorff distances for shape comparison. In *Eurographics Symposium on Point-Based Graphics* (eds Botsch, M. & Pajarola, R.) (Eurographics Association, 2007).
- Peyré, G., Cuturi, M. & Solomon, J. Gromov–Wasserstein averaging of kernel and distance matrices. In *Proc. 33rd International Conference on Machine Learning* (Journal of Machine Learning Research, 2016).
- Cuturi, M. Sinkhorn distances: lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26* (eds Burges, et al.) (NIPS, 2013).
- Petkova, M. D., Tkačik, G., Bialek, W., Wieschaus, E. F. & Gregor, T. Optimal decoding of cellular identities in a genetic network. *Cell* **176**, 844–855 (2019).
- Rodrigues, S. G. et al. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).
- Park, J. et al. Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science* **360**, 758–763 (2018).
- Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).
- Nüsslein-Volhard, C. & Wieschaus, E. Mutations affecting segment number and polarity in *Drosophila*. *Nature* **287**, 795–801 (1980).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

## Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

## Data pre-processing

For the cases for which normalized data was not available or used by the authors, we adopted the standard library size normalization in log-space, for example, if  $d_{ij}$  represents the raw count for gene  $i$  in cell  $j$ , we normalized it as

$$d_{ij} \rightarrow d'_{ij} = \log_2 \left( 10^5 \times \frac{d_{ij}}{\sum_k d_{kj}} + 1 \right).$$

Highly variable genes were identified by plotting the dispersion of a gene as a function of its mean and selecting the outliers above cut-off values (usually 0.125 for the mean and 1.5 for the dispersion).

In the Slide-seq datasets<sup>23</sup>, we summed up the transcriptomes of neighbouring cells by rounding the coordinates of the physical locations to the next integer multiple of 50. This resulted in a total of 8,331 (9,890) cells for the sagittal (coronal) section of the cerebellum. Low-quality locations were further filtered out by requiring at least 50 genes per cell, resulting in a total of 7,704 (8,258) for the sagittal (coronal) section. Marker genes for the reconstruction were randomly selected from the set of 747 genes. As one of the means of benchmarking the different reconstructions was to visually assess the expression pattern of *Pcp4*, we ensured that no genes with a Pearson correlation of  $|R| \geq 0.25$  with *Pcp4* were selected as marker genes.

## Mathematical formulation of novoSpaRc

The procedure used by novoSpaRc includes several steps. We first compute the graph-based distance matrices for  $N$  single cells in expression space,  $D^{\text{exp}} \in \mathbb{R}^{N \times N}$ , and for  $M$  locations,  $D^{\text{phys}} \in \mathbb{R}^{M \times M}$  (Extended Data Fig. 1a, Supplementary Methods). Then, optionally, if a reference atlas is available, we compute the matrix of disagreement,  $D^{\text{exp,phys}} \in \mathbb{R}^{N \times M}$ , between each of the cells to each of the locations, on the basis of the inverse correlation between the partial expression profile for each location given by the reference atlas and the respective expression profile for each cell. Equipped with these measures of intra- and inter-dataset distances, we set out to find an optimal (probabilistic) assignment of each of the single cells to cellular physical locations.

We formulate this problem as an optimization problem within the generalized framework of optimal transport<sup>14–16</sup>. Optimal transport is a mathematical framework that was first established in the eighteenth century by Gaspard Monge and was initially motivated by the question of the optimal (minimal cost) way to rearrange one pile of dirt into a different formation (the respective minimal cost is appropriately termed the ‘earth mover’s distance’). The framework evolved both theoretically and computationally<sup>15,16,21</sup> and was extended to the correspondence between pairwise similarity measures via the Gromov–Wasserstein distance<sup>19,20</sup>. Thus, in our context, it allows us to build on these results and tools to feasibly solve the cellular assignment problem.

We aim to find a probabilistic embedding,  $T \in \mathbb{R}_+^{N \times M}$ , of  $N$  single cells to  $M$  locations that would minimize the discrepancy between the pairwise graph-based distances in expression space and in physical space, and—if a reference atlas is available—simultaneously minimize the discrepancy between its values across the tissue and the expression profiles of embedded single cells. For each cell  $i$ , the value of  $T_{ij}$  is the relative probability of embedding it to location  $j$ . These optimization requirements over  $T$  are formulated as follows. We measure the pairwise discrepancy of  $T$  for the expression and physical spaces using the Gromov–Wasserstein discrepancy<sup>19</sup>

$$D_1(T) = \sum_{i,j,k,l} L(D_{i,k}^{\text{exp}}, D_{j,l}^{\text{phys}}) T_{i,j} T_{k,l},$$

where  $L$  is a loss function; specifically, we use the quadratic loss  $L(a, b) = \frac{1}{2} |a - b|^2$ . This term captures our preference to embed single cells such that their pairwise distance structure in expression space would resemble their pairwise distance structure in physical space. Intuitively, if expression profiles that correspond to cells  $i$  and  $k$  are embedded into cellular locations  $j$  and  $l$ , respectively, then the distance between  $i$  and  $k$  in expression space should correspond to the distance between  $j$  and  $l$  in physical space (for example, if  $i$  and  $k$  are close expression-wise they should be embedded into close locations, and vice versa). The discrepancy measure weighs these correspondences by the respective probability of the two embedding events.

To measure the match to existing prior knowledge, or an available reference atlas, we consider

$$D_2(T) = \sum_{i,j} D_{i,j}^{\text{exp,phys}} T_{i,j}.$$

This term represents the average discrepancy between cells and locations according to the reference atlas, weighted by  $T$ . Finally, we regularize  $T$  by favouring embeddings with higher entropy, where entropy is defined as

$$H(T) = - \sum_{i,j=1} T_{i,j} \log T_{i,j}$$

Intuitively, higher entropy implies more uncertainty in the mapping. Entropic regularization drives the solution away from arbitrary deterministic choices and was shown to be computationally efficient<sup>21</sup>.

Putting these together, we define the optimization problem for the optimal probabilistic embedding  $T^*$ :

$$T^* = \text{argmin} (1 - \alpha) D_1(T) + \alpha D_2(T) - \varepsilon H(T)$$

subject to

$$\sum_j T_{i,j} = p_i \quad \forall i \in \{1, \dots, N\}$$

$$\sum_i T_{i,j} = q_j \quad \forall j \in \{1, \dots, M\}$$

where  $\varepsilon$  is a non-negative regularization constant, and  $\alpha \in [0, 1]$  is a constant interpolating between the first two objectives, and can be set to  $\alpha = 0$  when no reference atlas is available. The constraints reflect the fact that the transport plan  $T$  should be consistent with the marginal distributions  $p \in \{p \in \mathbb{R}_+^N; \sum_i p_i = 1\}$  and  $q \in \{q \in \mathbb{R}_+^M; \sum_j q_j = 1\}$ , over the original input spaces of expression profiles and cellular locations, respectively.

These marginals can capture, for example, varying densities of single cells in the vicinity of different cellular grid locations, or the quality of different single-cell expression profiles (hence forcing low-quality single cells to have a smaller contribution to the reconstructed tissue-wide expression patterns). When such prior knowledge is lacking,  $p$  and  $q$  could be set to be uniform distributions.

We derive an efficient algorithm for this optimization problem, inspired by the combined results for entropically regularized optimal transport<sup>21</sup> and mapping based on Gromov–Wasserstein distance between metric-measure spaces<sup>20</sup> (Supplementary Methods).

Then, given the original single-cell expression profiles, represented by a matrix  $Y \in \mathbb{R}^{N \times g}$  (for  $N$  single cells and  $g$  genes), and the inferred probabilistic embedding  $T \in \mathbb{R}_+^{N \times M}$  (for  $N$  single cells and  $M$  locations), we can derive a virtual in situ hybridization (vISH),  $S = Y^T T \in \mathbb{R}_+^{g \times M}$

(for  $g$  genes and  $M$  locations), which contains the gene expression values for every cellular location of the target space.

Note again that because our mapping is probabilistic, each of the cellular locations of the vISH does not correspond to a single cell in the original data. Rather, the vISH represents the expression patterns over an averaged, stereotypical tissue from which the single cells could have originated.

### novoSpaRc algorithm

To spatially reconstruct gene expression, novoSpaRc performs the following steps:

1. Read the gene expression matrix.
  - 1a. Optional: select a random set of cells for the reconstruction;
  - 1b. Optional: select a small set of genes (for example, highly variable).
2. Construct the target space.
3. Set up the optimal-transport reconstruction.
  - 3a. Optional: use existing information of marker genes, if available.
4. Perform the spatial reconstruction including:
  - 4a. Assigning cells a probability distribution over the target space;
  - 4b. Deriving a vISH for all genes over the target space.

The novoSpaRc package, system requirements, installation guide and demo instructions are provided at <https://github.com/rajewsky-lab/novosparc>.

### Generating in silico single-cell data for the BDTNP dataset

To test the performance of novoSpaRc with single-cell resolution ground truth, we generated an in silico single-cell dataset for the BDTNP data<sup>13</sup>. In that case we have access to expression profiles for different locations across the embryo. We effectively dissociate the embryo by taking these expression profiles to be the expression profiles of single cells in our in silico set, masking their true original locations, and use novoSpaRc to reconstruct the original embryo (which may be done at lower spatial resolution).

### Identification of spatial archetypes

The identification of spatial archetypes is performed by clustering the spatial expression of a given set of genes. The gene expression is first clustered by hierarchical clustering at the vISH level, although in principle different clustering methods can be used. The number of archetypes is chosen by visually inspecting the resulting dendrogram. The expression values of each gene of the cluster are then averaged per location to produce the spatial archetype for that cluster. Representative genes for each cluster are identified by computing the Pearson correlation of each gene within the cluster against the spatial archetype. The derivation of the spatial archetypes strongly depends on the set of genes used. We observed that the set of highly variable genes generally resulted in sensible spatial archetypes. A list of genes that correspond to each archetype is provided in the Supplementary Information.

### Identification of zoned genes

For tissues with one-dimensional symmetry, we produce a ranking of highly zoned genes, both according to the original spatial expression patterns (Extended Data Fig. 2c, d) and the reconstructed patterns (Fig. 6a).

The input is a spatial expression matrix (either original or reconstructed), specifying the expression level of each gene in each of the spatial zones. Then, to find a ranked list of genes that are highly zoned towards the first or last spatial zones (for example, crypt in the liver), we first select all genes (i) whose highest expression occurs in that respective zone; (ii) whose maximum expression value is in the top 1% of all genes; and (iii) that are statistically significantly zoned. To compute the zonation significance of individual genes, we used a non-parametric test based on the Kendall's tau coefficient. The Kendall's tau coefficient

is a measure for the correspondence between two ranked lists—in our case, the expression values of a given gene over consecutive spatial zones and the numbering of the zones. Finally, the remaining genes are ranked according to their centre of mass.

The lists of predicted zoned genes based on novoSpaRc's reconstruction for the mammalian intestine and liver are available in the Supplementary Information.

### Gene ontology enrichment

We used GOrilla for gene ontology (GO) enrichment analysis<sup>25</sup>, in which GO enrichment was computed on the basis of target and background lists of genes (Supplementary Methods). For both the target and background lists of genes, we selected genes that had a maximum expression value in the top 10% of all genes. The target lists for genes that were zoned towards the boundaries of the one-dimensional spatial axes (crypt and V6 in intestine; layers 1 and 9 in liver) were further filtered to contain only genes that are statistically significantly zoned, as described in 'Identification of zoned genes'. The background lists contained the corresponding complements of the target lists.

### Identification of spatially informative genes

We use a spatial autocorrelation measure to rank genes as spatially informative. Specifically, we use Moran's  $I$  as a measure for global spatial autocorrelation. For each individual gene  $i$ , the Moran's  $I$  score for its spatial expression,  $y_i$ , over  $n$  cellular locations is:

$$I = \frac{n}{S_0} \frac{\sum_{i,j} z_i w_{ij} z_j}{\sum_i z_i^2}$$

where  $z_i = y_i - \bar{y}$ ,  $\bar{y}$  is the mean expression of gene  $i$ ,  $S_0 = \sum_{i,j} w_{ij}$  and  $w_{ij}$  is a spatial weights matrix, which we base on a  $k$ -nearest neighbours graph for each cellular location ( $k = 8$ ). To calculate the Moran's  $I$  score and the respective  $P$  values for different genes, we used the implementation of PySAL, a Python spatial analysis library<sup>27</sup>.

The Moran's  $I$  scores with their respective  $P$  values, based on novoSpaRc's reconstructions for all genes of the *Drosophila* embryo, zebrafish embryo and cerebellum, are available in the Supplementary Information.

### Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

### Data availability

The scRNA-seq datasets were acquired from the Gene Expression Omnibus (GEO) database with the following accession numbers: GSE99457 for the intestinal epithelium<sup>10</sup>, GSE84490 for the liver<sup>7</sup>, GSE95025 for the *Drosophila* embryo<sup>12</sup>, GSE66688 for the zebrafish embryo<sup>5</sup> and GSE107585 for the kidney<sup>24</sup>. The cerebellum Slide-seq datasets<sup>23</sup> were acquired from the Broad Institute Single Cell Portal ([https://portals.broadinstitute.org/single\\_cell/study/slide-seq-study](https://portals.broadinstitute.org/single_cell/study/slide-seq-study)). The individual *Drosophila* embryos dataset<sup>22</sup> is available as a supplementary information file of the original manuscript<sup>22</sup>. The BDTNP dataset was downloaded directly from the BDTNP webpage<sup>13</sup>.

### Code availability

A Python package for novoSpaRc, and the scripts for reconstructing selected tissues presented in the manuscript, are provided at <https://github.com/rajewsky-lab/novosparc>.

27. Rey, S. J. & Anselin, L. in *Handbook of Applied Spatial Analysis* (eds Fischer, M. & Getis, A.) 175–193 (Springer, 2010).

28. Tomancak, P. et al. Global analysis of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol.* **8**, R145 (2007).

# Article

**Acknowledgements** We thank A. Murray, A. Regev, T. Gregor, P. Rigollet, all members of our labs and many colleagues in the field for valuable comments and discussions. We thank L. Friedman for help with graphic design and illustration. This work was supported by the Israeli Science Foundation, through the I-CORE program (N.F.) and an Alexander von Humboldt Foundation Research Award (N.F.). N.K. was supported by grants DFG/GZ (Geschäftszeichen): RA 838/8-2 and DFG/GZ: KA 5006/1-1; and HGF Neurocore/GZ 0036-Phase 2-3. M.N. was supported by the James S. McDonnell Foundation, Schmidt Futures, the Israel Council for Higher Education and the John Harvard Distinguished Science Fellows Program within the FAS Division of Science of Harvard University. N.R. thanks Anna-Carina for useful discussions.

**Author contributions** N.R. conceived the structural correspondence assumption. N.K. and N.R. demonstrated the feasibility of such an assumption for spatial inference of toy models. M.N., N.K., N.F. and N.R. designed the research. M.N. developed the

optimal-transport-based spatial inference framework. M.N. and N.K. implemented the method and performed computational and data analyses. N.F. and N.R. supervised the study. All authors wrote the manuscript.

**Competing interests** The authors declare no competing interests.

## Additional information

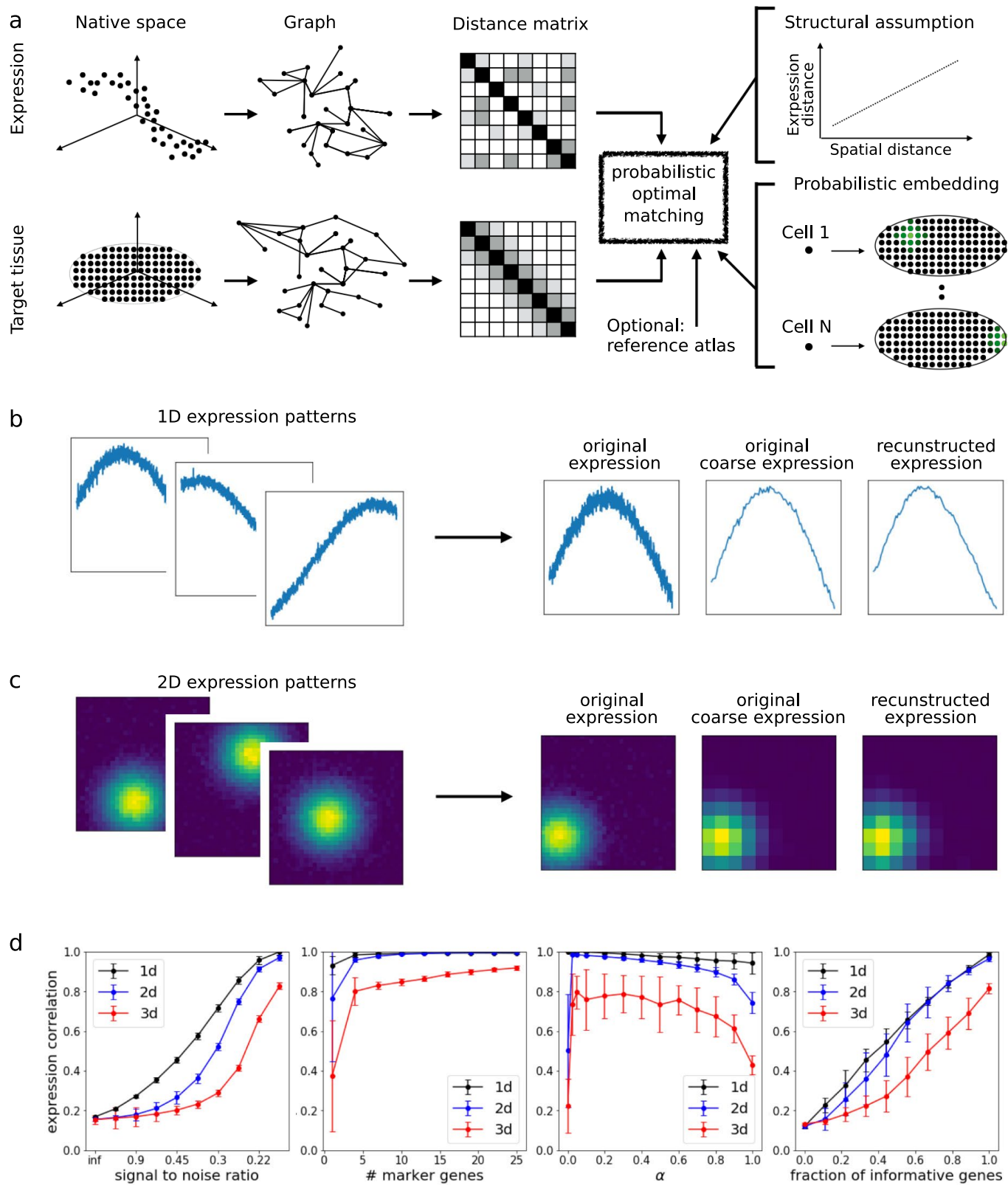
**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1773-3>.

**Correspondence and requests for materials** should be addressed to N.F. or N.R.

**Peer review information** *Nature* thanks Eileen Furlong and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

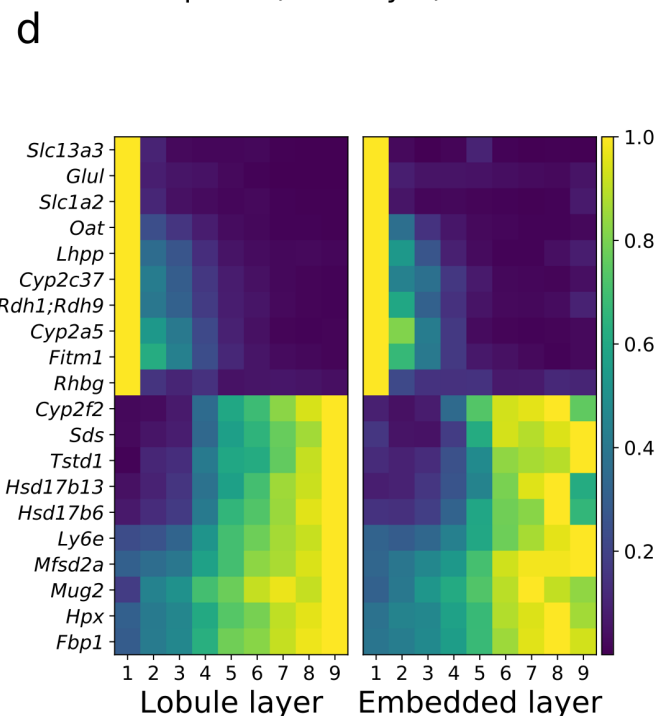
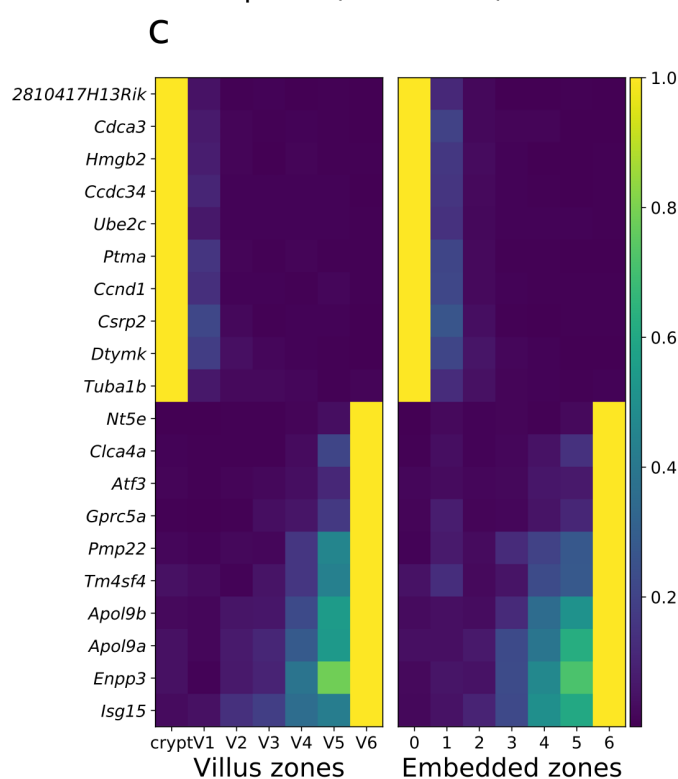
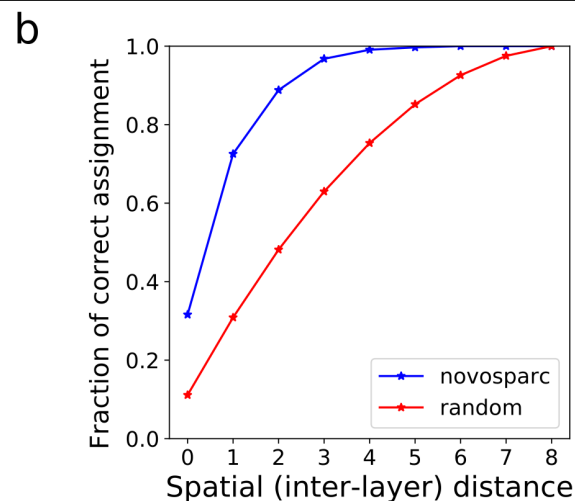
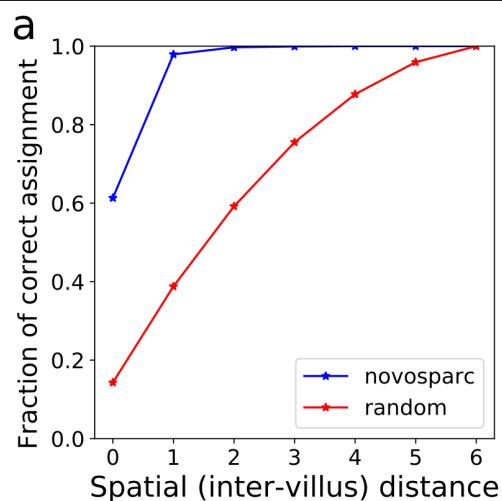




**Extended Data Fig. 1** | See next page for caption.

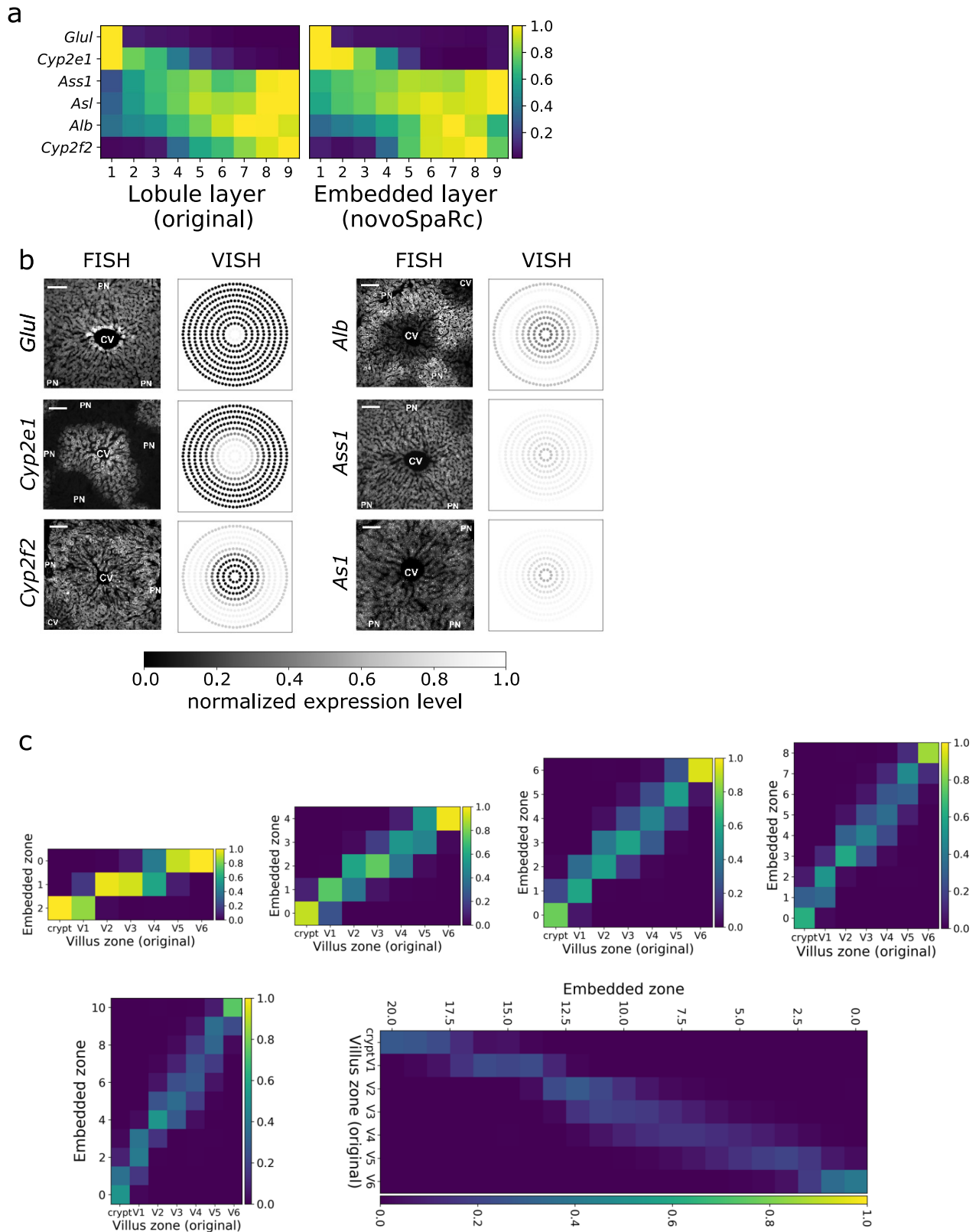
**Extended Data Fig. 1 | Overview of probabilistic optimal matching using novoSpaRc and corresponding generative model.** **a**, Based on the raw data of single cells in expression space and locations along a grid resembling the target tissue, graph structures are computed and distance matrices are derived from these graphs (Supplementary Methods). The two branches, and potentially a reference atlas, are aligned using novoSpaRc, under our structural correspondence assumption (distance in expression space on average monotonically increases with distance in physical space) and by using probabilistic embedding (Supplementary Methods). **b, c**, Left, visualization of noisy expression patterns for three random genes in models for 1-dimensional (1D) (**b**) and two-dimensional (2D) (**c**) tissues. Right, the original expression pattern for a representative gene, its coarse-grained representation

(decreased spatial resolution) and its reconstruction using novoSpaRc. **d**, The Pearson correlation of the reconstructed expression pattern data to the original synthetic expression data increases with increasing signal-to-noise ratio, with the number of marker genes and with the fraction of informative genes, and exhibits non-monotonic behaviour with the  $\alpha$  parameter. We note that  $\alpha$  is an interpolation parameter (defined in the Methods section 'Mathematical formulation of novoSpaRc') between using only a reference atlas ( $\alpha = 1$ ) and using only structural information (driven by the structural correspondence assumption) ( $\alpha = 0$ ). Results are averaged over 100 instantiations of the generative model; data are mean  $\pm$  s.d. The generative model and its default parameters are described in the Supplementary Methods.



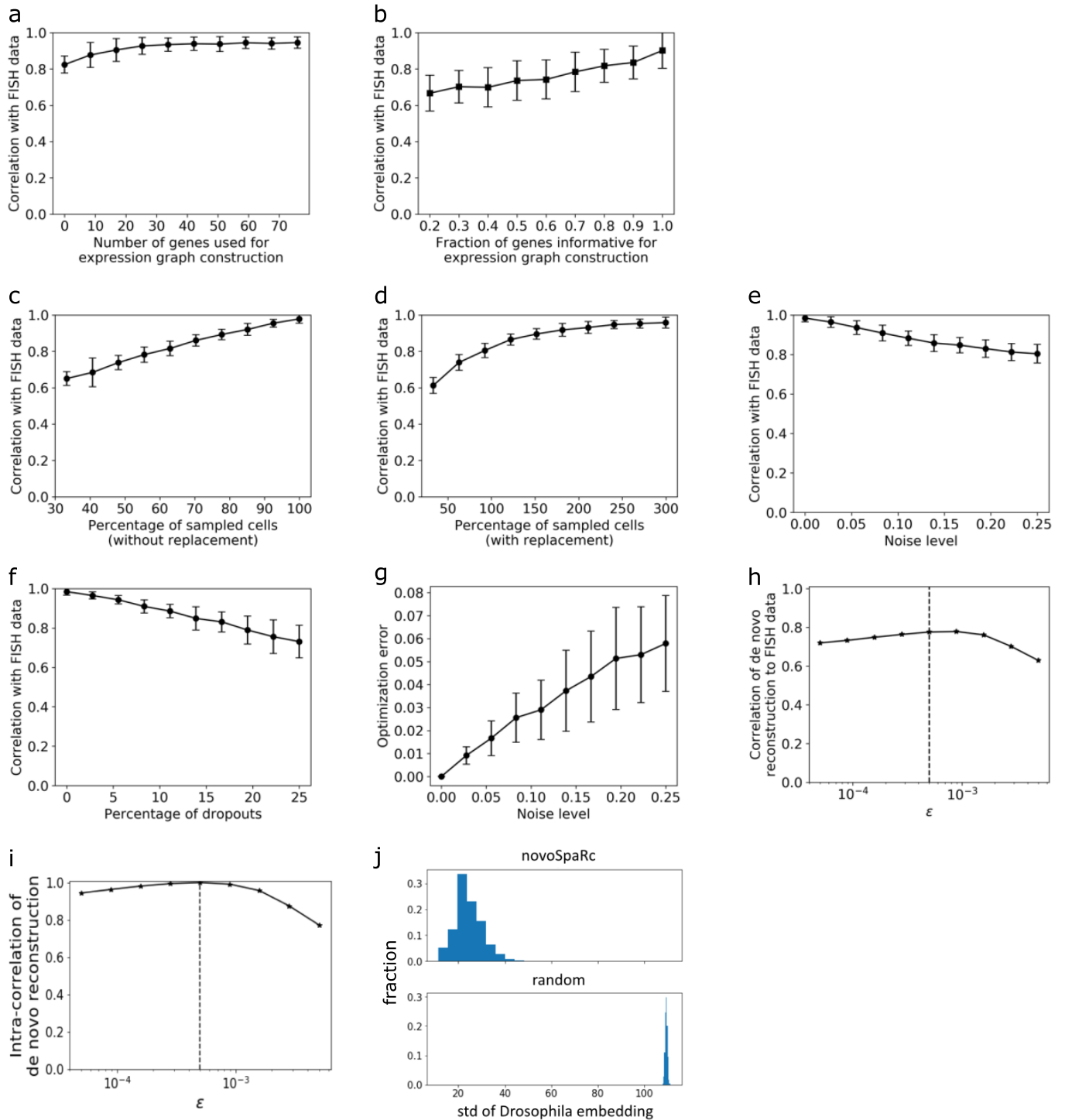
**Extended Data Fig. 2 | Evaluation of novoSpaRc reconstruction of the intestinal epithelium and the liver lobule. a, b,** The fraction of cells in the crypt-to-villus axis (a) and the liver lobule axis (b) that is correctly assigned to its corresponding original villus zone<sup>10</sup> and original lobule layer<sup>7</sup>, or is assigned to a zone up to  $d$  zones away from the original zone (x axis), is substantially higher than that of random assignment. **c, d,** novoSpaRc reconstructs the spatial expression patterns of the top zonated genes in the intestinal

epithelium (c) (10 top zonated genes towards the crypt, and 10 top zonated genes towards V6) and in the liver lobule (d) (10 top zonated genes towards the central vein (CV), and 10 top zonated genes towards the portal node (PN)). 2810417H13Rik is also known as *Pclaf*. The selection of the top zonated genes is described in the Methods. The expression level of each gene in c and d is normalized to its maximum value.



**Extended Data Fig. 3 | novoSpaRc reconstruction of the intestinal epithelium and the liver lobule is robust and consistent with changing grid resolution. a, b.** Examples of FISH expression patterns of six zoned genes across the liver lobules, comparing the reconstructed (de novo vISH data) expression patterns produced by novoSpaRc to the expression patterns reported in a previous study<sup>7</sup> (a), and the original (FISH) data (adapted from the same study<sup>7</sup>) (b). The visualization in a is a heat map, which shows the expression values of each gene across the lobule layers. The visualization of the

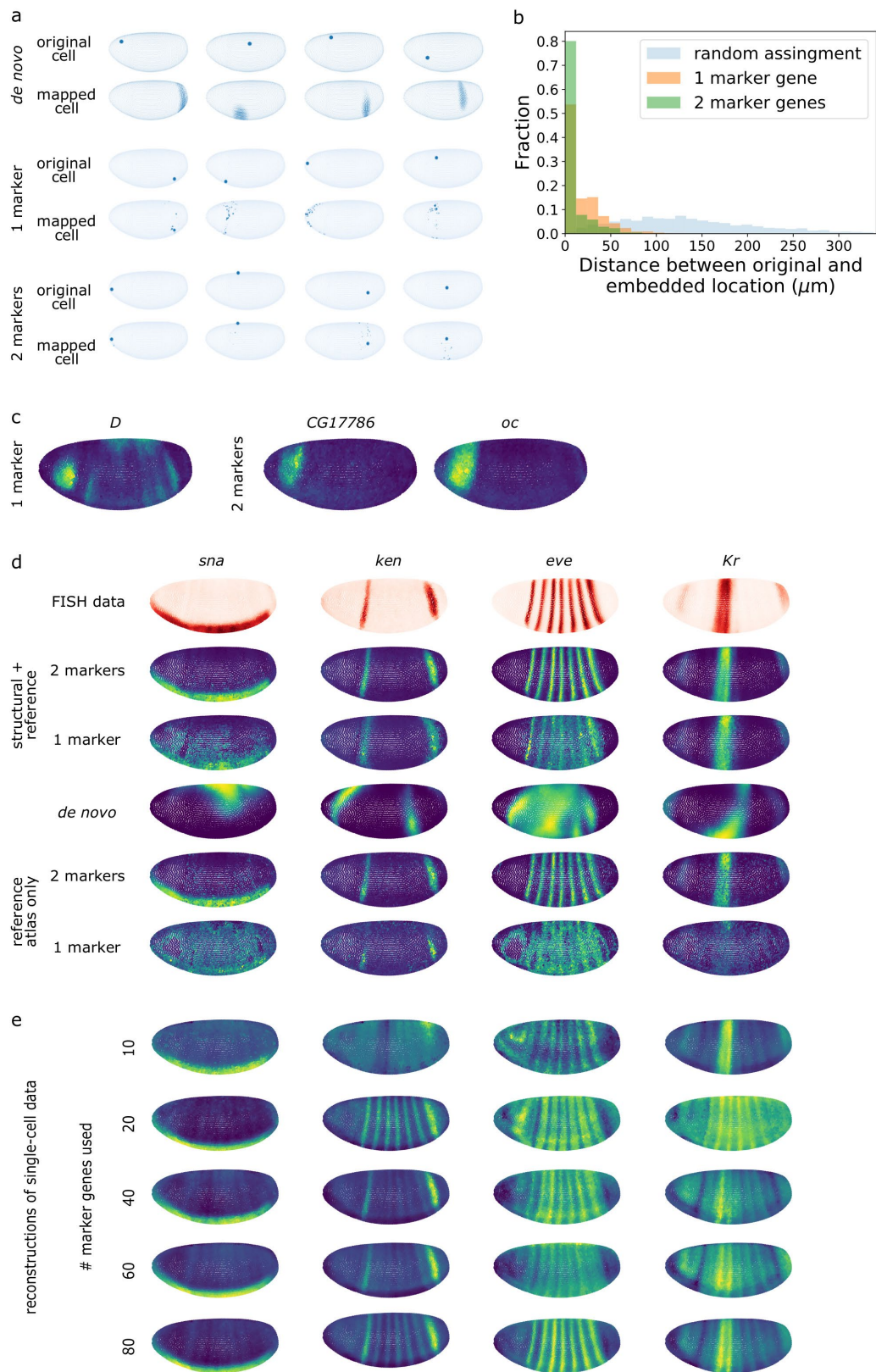
reconstructed vISH data in b is intended to be comparable to the FISH images, and therefore the 1D reconstructed coordinates are projected onto a polar coordinate system (central vein–middle, portal node–outer circumference). c, The successful de novo reconstruction of the intestinal epithelium dataset<sup>9</sup> is achieved for varying numbers of layers used for the target space (including both lower and higher numbers of layers compared with the original number (seven) of reference layers). The expression level of each gene is normalized to its maximum value.



**Extended Data Fig. 4 | novoSpaRc reconstruction of the *Drosophila* embryo on the basis of the BDTNP dataset is robust and self-consistent. **a, b**, The Pearson correlation of the reconstructed expression patterns to the original FISH expression data<sup>12</sup> increases with the number of genes used to construct the structural cellular graph in expression space (**a**), and with the fraction of those genes that are spatially informative (**b**). Spatially non-informative genes in this case were simulated as random Gaussian variables with mean and s.d. comparable to that of the original set of genes. **c–f**, The Pearson correlation of the reconstructed expression patterns to the original FISH expression data<sup>12</sup> increases with the percentage of sampled single cells (without replacement) (**c**) and with the percentage of sampled single cells (with replacement) (**d**), and steadily decreases with noise level (**e**) and with the percentage of dropouts in the data (**f**). **g**, The mean value and variance of the optimization objective**

function (which we aim to minimize) increase with noise level. The results in **a–g** are averaged over 100 random choices of two marker genes; data are mean  $\pm$  s.d. **h**, The Pearson correlation of the de novo reconstructed expression patterns to the original FISH data varies gradually with the entropic regularization parameter  $\epsilon$ . **i**, The Pearson correlation of embedded de novo expression patterns of the BDTNP dataset<sup>12</sup> for different values of the entropic regularization parameter  $\epsilon$  with the expression pattern for  $\epsilon = 5 \times 10^{-5}$  (vertical dotted line). **j**, The spatial s.d. of embedded cells over the *Drosophila* embryo of the BDTNP dataset derived from de novo reconstruction by novoSpaRc is significantly lower than the s.d. derived from randomized embedding ( $P < 10^{-200}$ , two-sided Kolmogorov–Smirnov test). Histograms show results for all 3,039 cells.

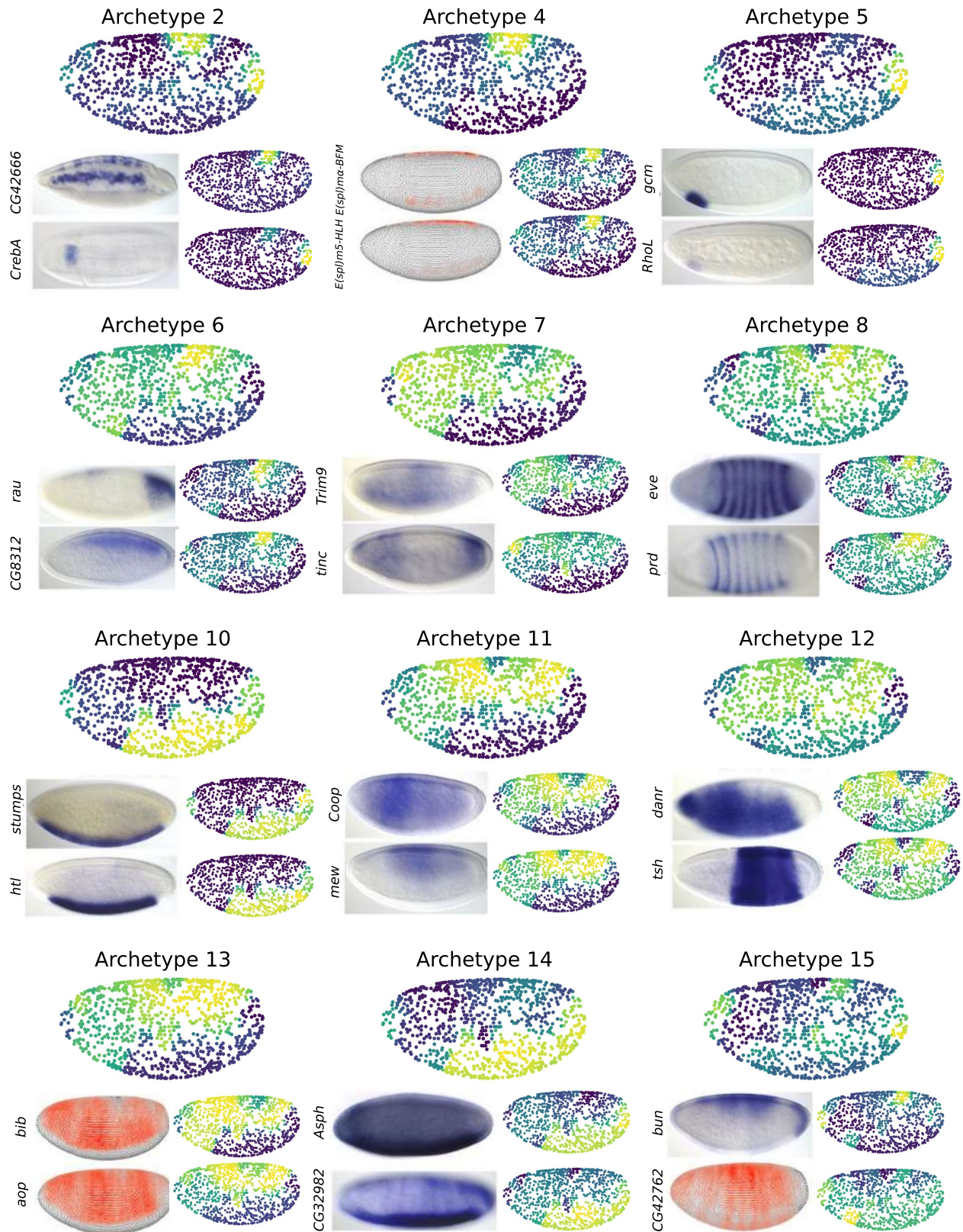




**Extended Data Fig. 5** | See next page for caption.

**Extended Data Fig. 5 | novoSpaRc accurately reconstructs the *Drosophila* embryo on the basis of the BDTNP dataset and single-cell data.** **a**, Examples of mapping probabilities of single cells produced by novoSpaRc for the *Drosophila* embryo, using the BDTNP dataset<sup>13</sup>. The predicted spatial positions of cells are distributed over relatively many locations when reconstruction is done de novo, and are more localized when marker genes are used. **b**, Histogram of Euclidean distances between the original cellular location of single cells and the most likely location predicted by novoSpaRc using one and two marker genes, compared to a histogram for random spatial predictions. **c**, The expression patterns of the two marker genes and one marker gene that were used for the results presented in **a**, **b** and in Fig. 3d, e. **d**, Visualization of

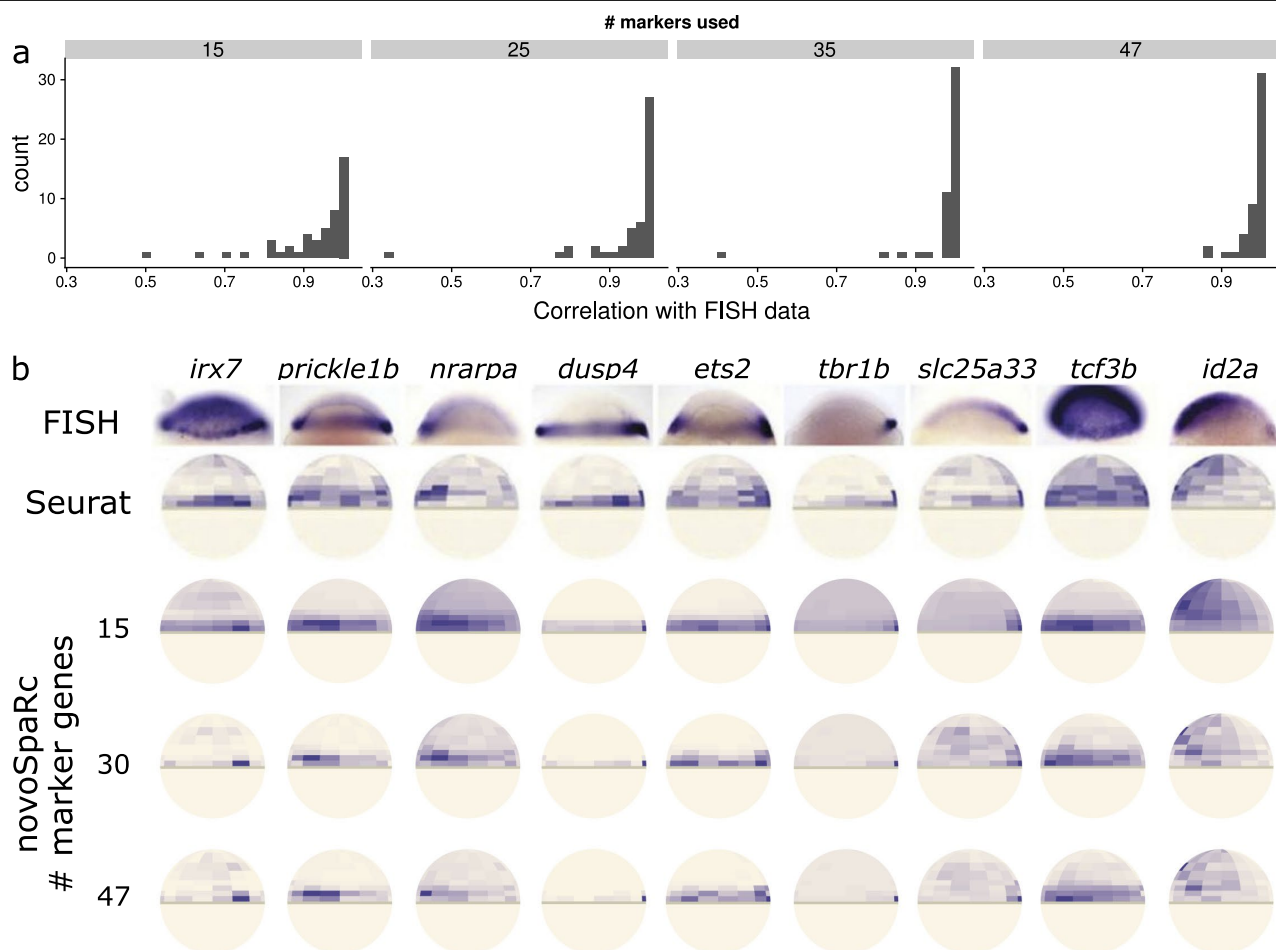
reconstruction results for four transcription factors. The original FISH data are compared to reconstruction by novoSpaRc that exploits both structural and marker gene information (using two marker genes and one marker gene), and reconstruction without any marker gene information (de novo). Reconstruction that uses both structural and marker gene information (or a reference atlas) outperforms reconstruction that is based solely on a reference atlas. **e**, Visualization of novoSpaRc-based reconstruction results for the four transcription factors, based on single-cell data<sup>12</sup> that exploit both structural and marker gene information (using 10–80 marker genes). The results in **a–d** are based on the BDTNP dataset<sup>13</sup>, and the results in **e** are based on a single-cell dataset<sup>12</sup>.



**Extended Data Fig. 6 | novoSpaRc identifies spatially informative archetypes by using scRNA-seq data for the *Drosophila* embryo.** The archetypes shown complement those of Fig. 4c, d. Preferred spatial positioning is denoted by colouring ranging from blue (low) to yellow (high). FISH images were taken from the BDGP database<sup>28</sup>. For genes for which an image was not available, DVEX<sup>12</sup> was used instead. Two representative genes are

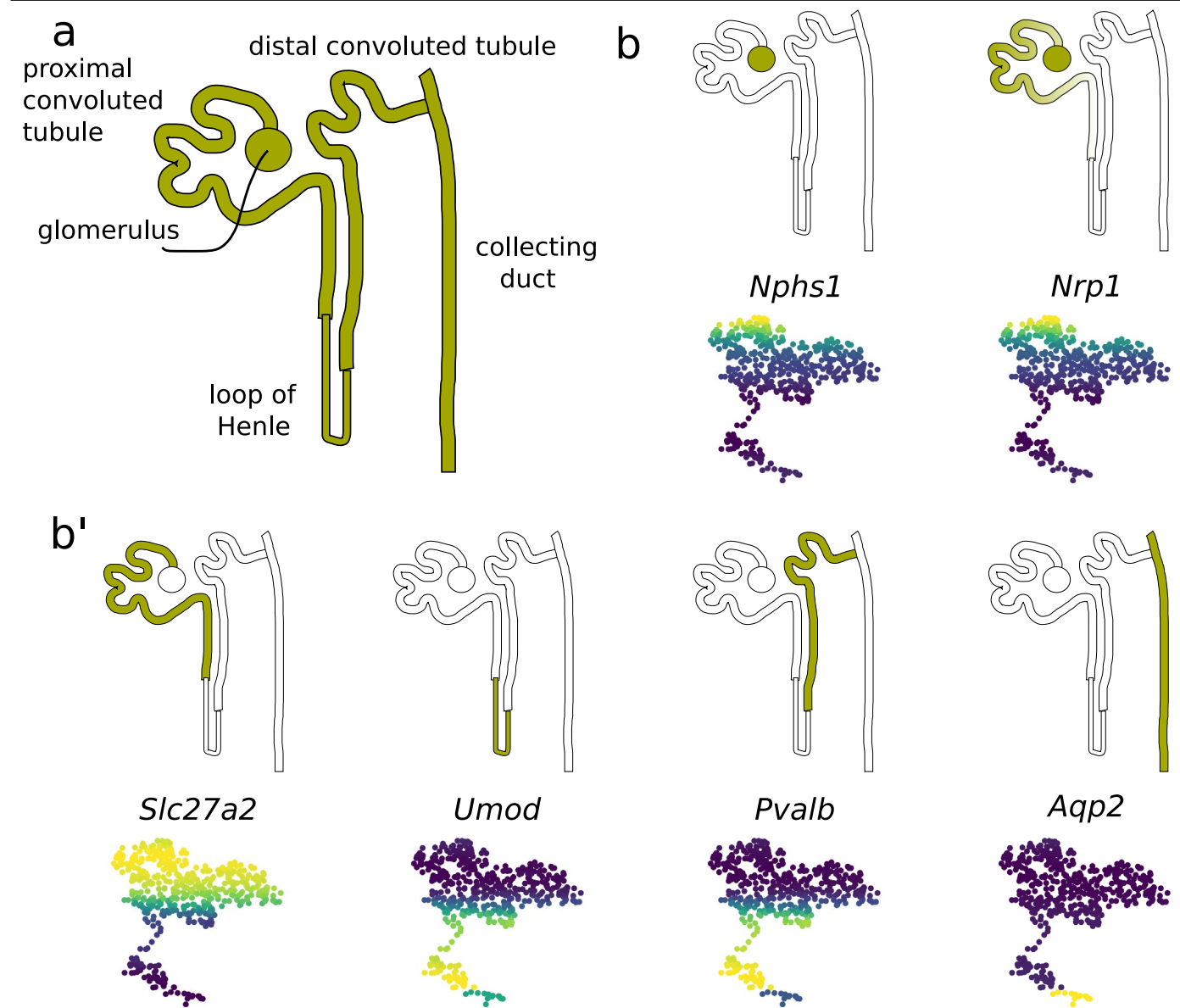
shown for each spatial archetype. novoSpaRc accurately groups genes expressed in a particular domain—for example, the subdomain of the mesoderm, which is characterized by the transcription factor *gcm* (Archetype 5)—whereas it does not capture the details of the fine expression patterns of pair-rule genes (Archetype 8). *CG42666* is also known as *prage*.





**Extended Data Fig. 7 | novoSpaRc reconstructs the zebrafish embryo.**  
**a**, Histograms assessing the increase in the accuracy of novoSpaRc reconstruction (measured by the Pearson correlation with FISH data<sup>5</sup>) with increasing number of marker genes. **b**, novoSpaRc reconstructs patterns of gene expression in the zebrafish embryo on the basis of only 15 marker genes,

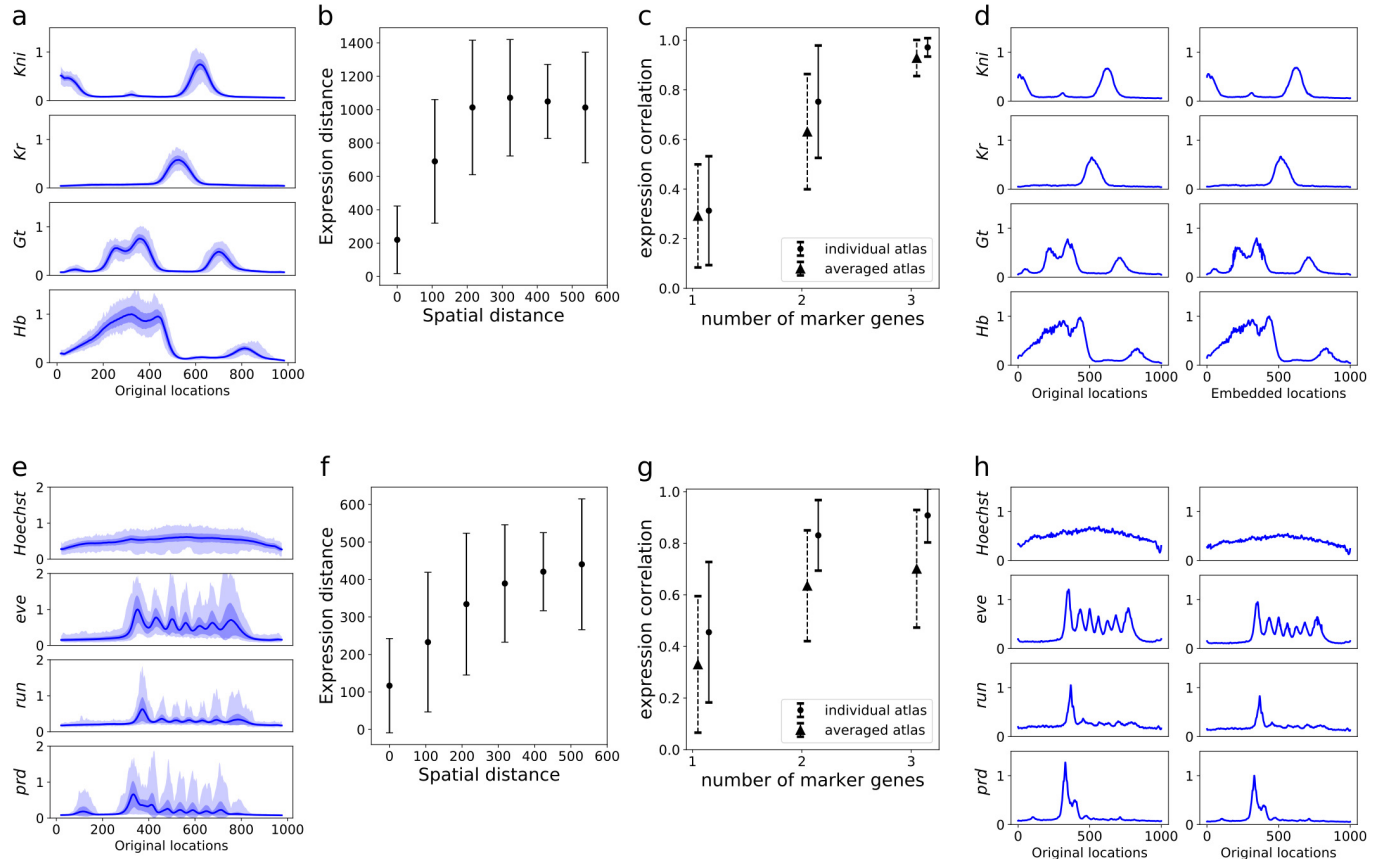
and the results improve as the number of marker genes increases. Top row, FISH data (reproduced from ref. <sup>5</sup>); second row: Seurat predictions using 47 marker genes<sup>5</sup>; bottom three rows: novoSpaRc predictions using 15, 30 and 47 marker genes. The genes shown were not used in any of the reconstructions.



**Extended Data Fig. 8 | novoSpaRc reconstructs a whole-kidney dataset de novo.** **a**, Sketch of the major cell types that are reconstructed with novoSpaRc. **b**, Representative marker genes for each of the cell types shown in **a**. Top rows depict a rough positioning for each cell type in yellow-green;

bottom rows show the gene expression predicted by novoSpaRc in the reconstructed tissue. *Nphs1*, podocytes; *Nrpl*, endothelial cells; *Slc27a2*, proximal tubule cells; *Umod*, loop of Henle; *Pvalb*, distal convoluted tubules; *Aqp2*, collecting duct cells. Expression ranges from low (blue) to high (yellow).

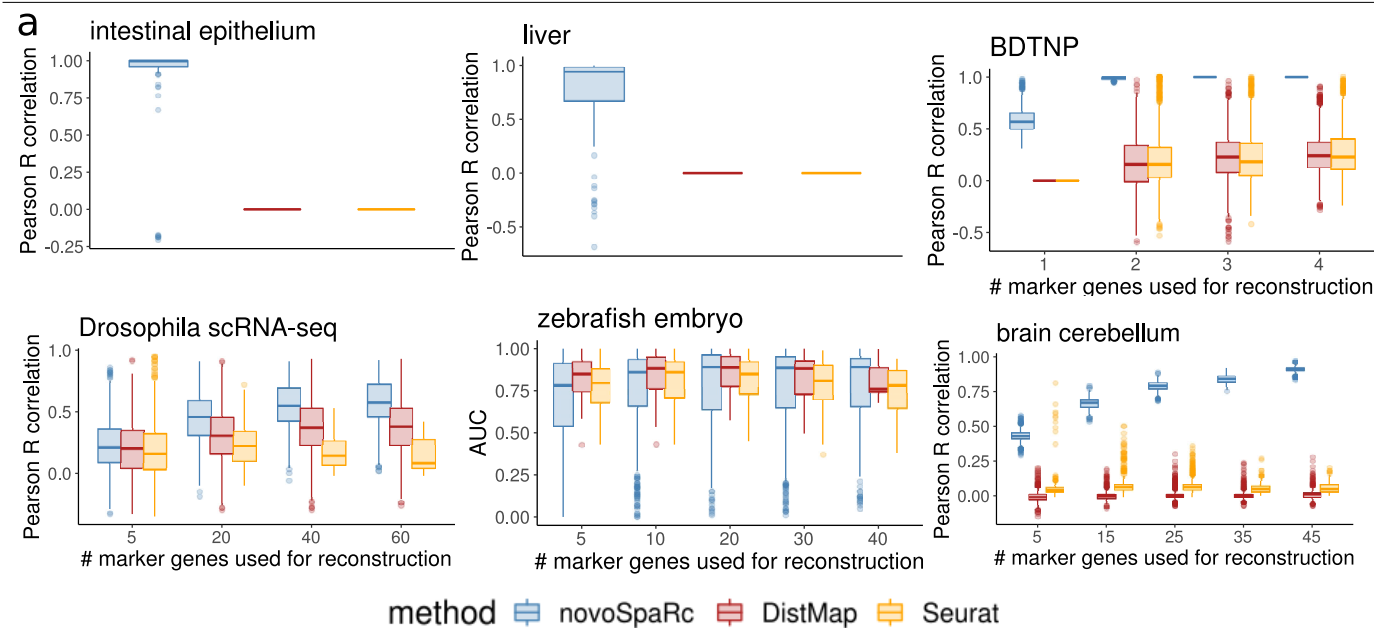




**Extended Data Fig. 9 | NovoSpaRc reconstructs single *Drosophila* embryos.**

**a, e**, The averaged original expression of four gap genes (**a**) and four pair-rule genes (**e**) is shown for 101 and 177 individual *Drosophila* embryos, respectively<sup>22</sup>. Solid line, mean; dark shadow, s.d.; light shadow, minimum and maximum values over all embryos. **b, f**, Demonstration of the monotonic relationship between cellular pairwise distances in expression and physical space, consistent with the structural correspondence assumption. Data are mean  $\pm$  s.d. **c, g**, The Pearson correlation increases with the number of marker genes used by novoSpaRc for the reconstruction of the remaining genes ( $\alpha = 0.5$ ) for both gap genes (**c**) and pair-rule genes (**g**). Using a reference atlas that corresponds to the individual embryo being reconstructed ('individual

atlas') results in a consistently higher reconstruction quality than using an averaged reference atlas over all embryos ('averaged atlas'). Data are mean  $\pm$  s.d. **d, h**, Examples of the reconstruction of the expression patterns across a single random embryo, in which the reconstruction of each of the four genes is performed using the three complement genes as a reference, for both gap genes (**d**) and pair-rule genes (**h**). Note that the reconstructed expression patterns presented in **d, h** were computed while the corresponding gene in each case was not used for the reconstruction. The expression level of each gene in **a, d, e, h** is normalized to the maximum value over the mean expression of all embryos.

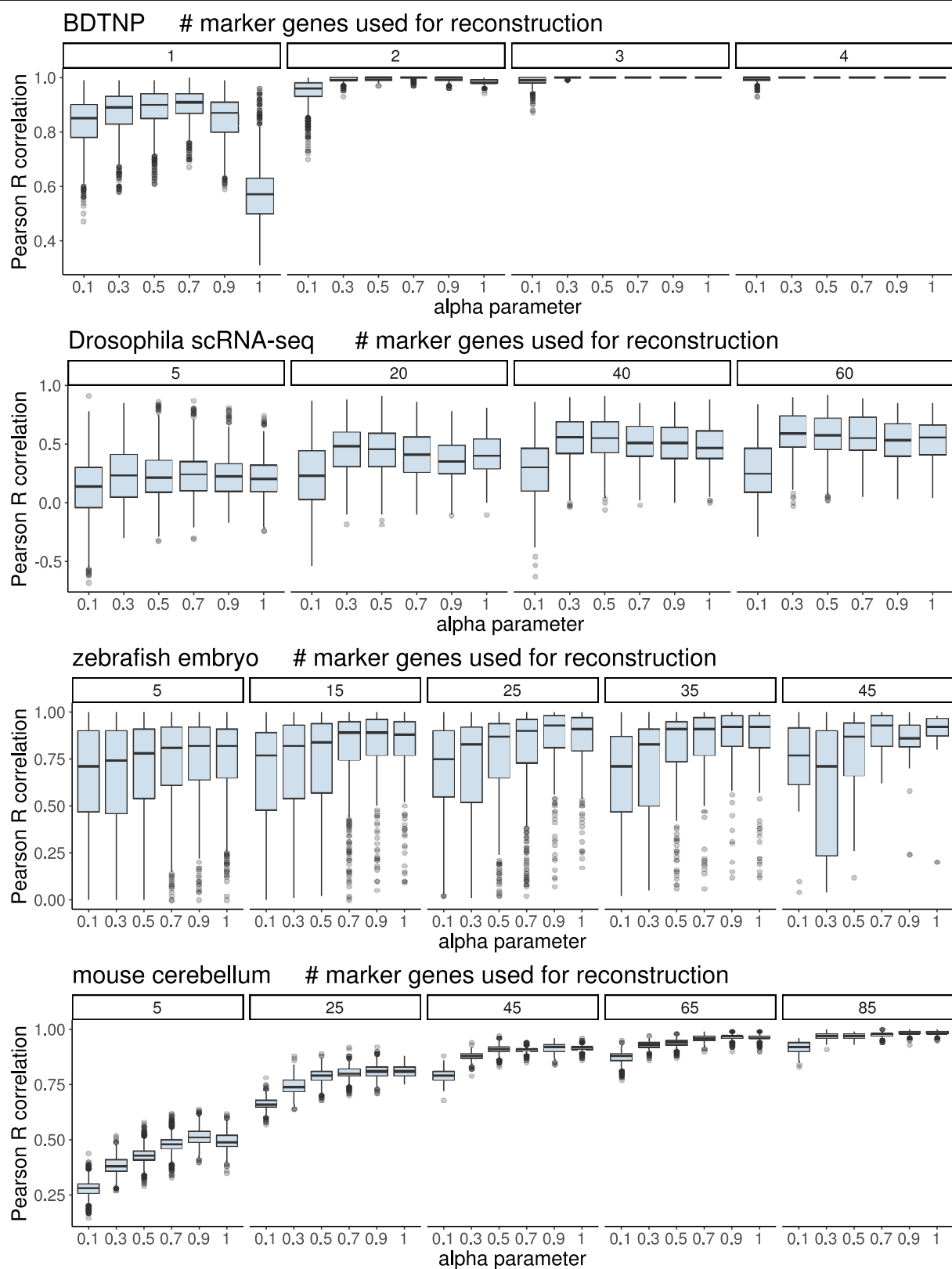


**b**

	Seurat	DistMap	novoSpaRc
Spatial mapping with reference atlas	✓	✓	✓
Reference atlas can have continuous values	X	X	✓
Spatial mapping <i>de novo</i>	X	X	✓
Does not require predetermined shape	✓	✓	X
Can exploit structural information	X	X	✓
Can use continuous expression data	X	X	✓
Can be applied to complex tissues	X	✓	✓
Does not require data imputation	X	✓	✓
Does not require a threshold	✓	X	✓

**Extended Data Fig. 10 | Comparison of spatial reconstruction with novoSpaRc versus available methods that fully rely on a reference atlas.** **a**, The Pearson correlation of the predicted versus the original spatial gene expression is shown as a function of the top 100 highly variable genes for the intestinal epithelium and liver datasets, or the number of marker genes used for the reconstruction for the BDTNP dataset, the *Drosophila* and zebrafish embryos and the brain cerebellum (84, 84, 45 and 745 genes, respectively). For the 1D datasets, the reconstructions are done *de novo* (with no reference atlas) and the existing baseline methods are inapplicable. For the liver, the last lobule layer was removed from the analysis, as only five cells were associated with it. For the 2D datasets, correlations are computed only for genes that were not

used for the reconstructions. Note that for the *Drosophila* embryo novoSpaRc outperforms DistMap<sup>12</sup>, and for the zebrafish embryo novoSpaRc performs comparably to or better than Seurat<sup>5</sup>—although those methods were developed and tailored for the *Drosophila* and zebrafish embryos, respectively, and the best-performing threshold was chosen for DistMap. For the box plots, the centre line is the median, box limits are the 0.25 and 0.75 quantiles and whiskers extend to  $\pm 2.698$  s.d. For the BDTNP dataset, the *Drosophila* and zebrafish embryos and the brain cerebellum, the results are shown for 100 random choices of marker genes. **b**, The intrinsic characteristics of novoSpaRc compared against Seurat<sup>5</sup> and DistMap<sup>12</sup>.



**Extended Data Fig. 11 | Reconstruction quality varies with the  $\alpha$  parameter.** Reconstructions of the BDTNP dataset, the *Drosophila* and zebrafish embryos and the brain cerebellum, with varying numbers of marker genes used for the reconstruction and different values of the  $\alpha$  parameter. The reconstruction quality is quantified by calculating Pearson correlations between the predicted and the original patterns of gene expression for all genes that were not used as markers for the reconstruction. The quality of the reconstruction decreases for  $\alpha=1$  in the BDTNP and brain cerebellum cases, which corresponds to

reconstructing based only on reference marker genes, without taking the structural correspondence assumption into account. We note that  $\alpha$  is an interpolation parameter (defined in the Methods section 'Mathematical formulation of novoSpaRc') between using only a reference atlas ( $\alpha=1$ ) and using only structural information (driven by the structural correspondence assumption) ( $\alpha=0$ ). For the box plots, the centre line is the median, box limits are the 0.25 and 0.75 quantiles and whiskers extend to  $\pm 2.698$  s.d. Results are shown for 100 random choices of marker genes.

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☒ ☐ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☒ ☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection

No software was used for data collection. All data shown in the manuscript is already publicly available.

Data analysis

We wrote custom software code which is available online on Github (distributed under the MIT License, version 0.2.2, <https://github.com/rajewsky-lab/novosparc>). The code is written in python and uses commonly used python libraries (numpy, matplotlib, sklearn, scipy, ot). To calculate spatial autocorrelation we used the implementation of PySAL (version 2.0.0), a Python spatial analysis library. We also used an implementation of the Gromov-Wasserstein transport method by Erwan Vautier (distributed under the MIT License).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

No datasets were generated during the current study. The single cell datasets analyzed for the current study were acquired from the GEO database with the following GEO accession numbers: GSE99457 for the intestinal epithelium, GSE84490 for the liver, GSE95025 for the Drosophila embryo, GSE66688 for the zebrafish embryo and GSE107585 for the kidney. The cerebellum Slide-seq datasets were acquired from the Broad Institute Single Cell Portal ([https://portals.broadinstitute.org/single\\_cell/study/slide-seq-study](https://portals.broadinstitute.org/single_cell/study/slide-seq-study)). The individual Drosophila embryos dataset (Petkova, M.D., et al., Cell 2019) is available as Supplemental Information files of the original manuscript. The BDTNP dataset was downloaded directly from the BDTNP webpage (<http://bdtnp.lbl.gov>).

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences    ☐ Behavioural & social sciences    ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes for Fig. 3c & Ext. Data Figs. 2c 6a-g, 11b,c, 17,18 were based on 100 instantiations, and for Figs. 2b,f, 3b, 5b & Ext. Data Figs. 10b, 16b,c,f,g there was no subsampling of the data.
Data exclusions	No data were excluded
Replication	Experimental replication was not attempted and is not applicable to this study
Randomization	Since the single cell transcriptomes are unique and technically not reproducible, randomization was not applicable to the study
Blinding	No datasets were generated during the current study, and therefore blinding was not applicable.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging



# PGRMC2 is an intracellular haem chaperone critical for adipocyte function

<https://doi.org/10.1038/s41586-019-1774-2>

Received: 29 November 2018

Accepted: 1 October 2019

Published online: 20 November 2019

Andrea Galmozzi<sup>1</sup>, Bernard P. Kok<sup>1</sup>, Arthur S. Kim<sup>1</sup>, J. Rafael Montenegro-Burke<sup>2</sup>, Jae Y. Lee<sup>1</sup>, Roberto Spreafico<sup>3</sup>, Sarah Mosure<sup>4,5</sup>, Verena Albert<sup>1</sup>, Rigo Cintron-Colon<sup>1</sup>, Cristina Godio<sup>1</sup>, William R. Webb<sup>2</sup>, Bruno Conti<sup>1</sup>, Laura A. Solt<sup>4</sup>, Douglas Kojetian<sup>5</sup>, Christopher G. Parker<sup>6,7</sup>, John J. Peluso<sup>8</sup>, James K. Pru<sup>9</sup>, Gary Siuzdak<sup>2,7</sup>, Benjamin F. Cravatt<sup>7</sup> & Enrique Saez<sup>1\*</sup>

Haem is an essential prosthetic group of numerous proteins and a central signalling molecule in many physiologic processes<sup>1,2</sup>. The chemical reactivity of haem means that a network of intracellular chaperone proteins is required to avert the cytotoxic effects of free haem, but the constituents of such trafficking pathways are unknown<sup>3,4</sup>. Haem synthesis is completed in mitochondria, with ferrochelatase adding iron to protoporphyrin IX. How this vital but highly reactive metabolite is delivered from mitochondria to haemoproteins throughout the cell remains poorly defined<sup>3,4</sup>. Here we show that progesterone receptor membrane component 2 (PGRMC2) is required for delivery of labile, or signalling haem, to the nucleus. Deletion of PGRMC2 in brown fat, which has a high demand for haem, reduced labile haem in the nucleus and increased stability of the haem-responsive transcriptional repressors Rev-Erb $\alpha$  and BACH1. Ensuing alterations in gene expression caused severe mitochondrial defects that rendered adipose-specific PGRMC2-null mice unable to activate adaptive thermogenesis and prone to greater metabolic deterioration when fed a high-fat diet. By contrast, obese-diabetic mice treated with a small-molecule PGRMC2 activator showed substantial improvement of diabetic features. These studies uncover a role for PGRMC2 in intracellular haem transport, reveal the influence of adipose tissue haem dynamics on physiology and suggest that modulation of PGRMC2 may revert obesity-linked defects in adipocytes.

A small molecule has recently been isolated that stimulated adipogenesis<sup>5</sup> by acting as a gain-of-function ligand for PGRMC2, a poorly characterized single-pass transmembrane protein localized in the endoplasmic reticulum and the nuclear envelope<sup>5–8</sup>. PGRMC2 belongs to the membrane-associated progesterone receptor (MAPR) family, the members of which share a non-covalent haem-binding domain<sup>9</sup>. Other MAPR proteins (such as PGRMC1, neudesin and neoferricin) bind haem reversibly<sup>9</sup>. We found that PGRMC2 also reversibly bound haem<sup>5</sup>. Of note, addition of haem boosts adipogenesis, whereas inhibition of biosynthesis blocks differentiation<sup>10</sup>. The adipogenic effects of haem have been linked to the nuclear receptor Rev-Erb $\alpha$ <sup>11,12</sup>, a transcriptional repressor with a dual role in adipogenesis: it is required early on, but it must be degraded for differentiation to proceed<sup>13</sup>. Haem is a ligand for Rev-Erb $\alpha$ <sup>14,15</sup>, and binding of haem leads to eventual Rev-Erb $\alpha$  degradation. Notably, the adipogenic effect of the PGRMC2 activator was dependent on Rev-Erb $\alpha$  signalling<sup>5</sup>, hinting that PGRMC2 activation may stimulate adipogenesis by facilitating haem delivery to the nucleus to induce Rev-Erb $\alpha$  degradation. Here, we have examined a role for PGRMC2 in intracellular haem mobilization.

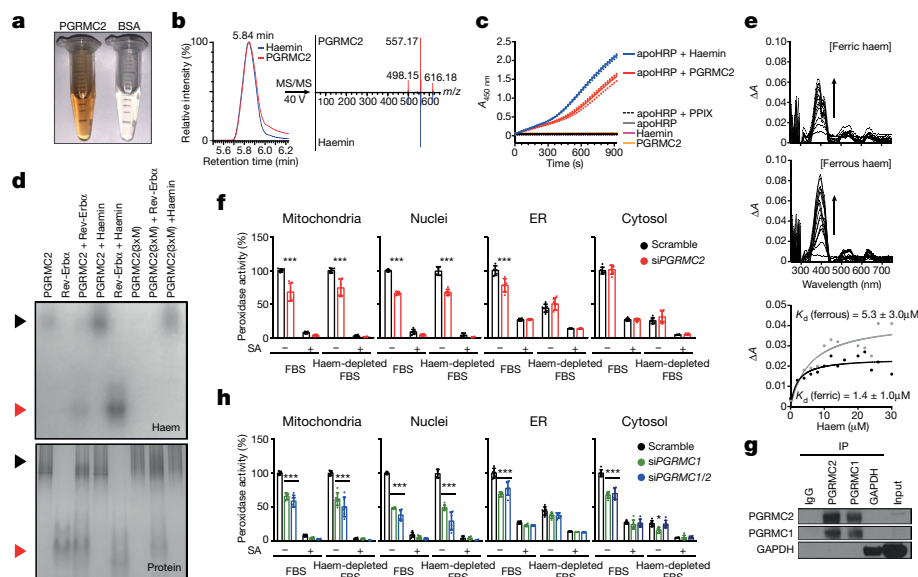
## PGRMC2 traffics mitochondrial haem

PGRMC2 protein purified from *Escherichia coli* was noticeably reddish in colour (Fig. 1a). Its spectrum revealed the Soret peak of haemoproteins at 390–430 nm (Extended Data Fig. 1a), and liquid chromatography–mass spectrometry (LC–MS/MS) showed a 616.18-Da peak, corresponding to iron–protoporphyrin IX (Fig. 1b, Extended Data Fig. 1b, c), confirming that PGRMC2 co-purified with haem. To test the ability of PGRMC2 to transfer haem (a requirement for a haem-mobilizing chaperone), we incubated PGRMC2 with apohorseradish peroxidase (apo-HRP), an inactive form of the enzyme lacking its prosthetic haem. Incubation of apo-HRP with haemin or PGRMC2 increased HRP activity, reflecting conversion of apo-HRP into active, haem-bound holoHRP (Fig. 1c)—thus indicating that PGRMC2 can transfer haem to other proteins. To test the ability of PGRMC2 to transfer haem to Rev-Erb $\alpha$  itself, apo-Rev-Erb $\alpha$  was incubated with PGRMC2, the mixture was separated by native electrophoresis, and the gel was stained for haem and protein. In-gel staining revealed haem bound to PGRMC2, but not to apo-Rev-Erb $\alpha$  (Fig. 1d). By contrast, apo-Rev-Erb $\alpha$  incubated with wild-type PGRMC2, but not with a haem-binding

<sup>1</sup>Department of Molecular Medicine, The Scripps Research Institute, La Jolla, CA, USA. <sup>2</sup>Scripps Center for Metabolomics, The Scripps Research Institute, La Jolla, CA, USA. <sup>3</sup>Institute for Quantitative and Computational Biology, University of California, Los Angeles, CA, USA. <sup>4</sup>Department of Immunology and Microbiology, The Scripps Research Institute, Jupiter, FL, USA.

<sup>5</sup>Department of Integrative Structural and Computational Biology, The Scripps Research Institute, Jupiter, FL, USA. <sup>6</sup>Department of Chemistry, The Scripps Research Institute, Jupiter, FL, USA.

<sup>7</sup>Department of Chemistry, The Scripps Research Institute, La Jolla, CA, USA. <sup>8</sup>Department of Cell Biology, University of Connecticut Health Center, Farmington, CT, USA. <sup>9</sup>Center for Reproductive Biology, Department of Animal Sciences, Washington State University, Pullman, WA, USA. \*e-mail: esaez@scripps.edu



**Fig. 1 | PGRMC2 controls the intracellular distribution of labile haem.**

**a**, Purified PGRMC2 is similar in colour to haemoproteins. **b**, LC–MS/MS spectra of PGRMC2 and haemin standard. **c**, Peroxidase activity of apo-HRP with PGRMC2. PGRMC2, haemin, apo-HRP and apo-HRP plus protoporphyrin IX (PPIX) show no activity. Haemin served as positive control. Technical duplicates are shown. **d**, Native PAGE of wild-type and haem-binding mutant (3×M) PGRMC2 and apo-Rev-Erb $\alpha$  ligand-binding domain (LBD) alone or in combination stained in-gel for haem (top) or protein (bottom). Black arrows, PGRMC2; red arrows, Rev-Erb $\alpha$ . Haemin (20  $\mu$ M) served as a positive control. PGRMC2 3×M and apo-Rev-Erb $\alpha$  LBD show no haem staining. **e**, Differential spectroscopy of PGRMC2 haem-binding domain with increasing amounts of ferric or ferrous (in presence of 10 mM dithionite) haemin. Titration curves represent differential absorbance

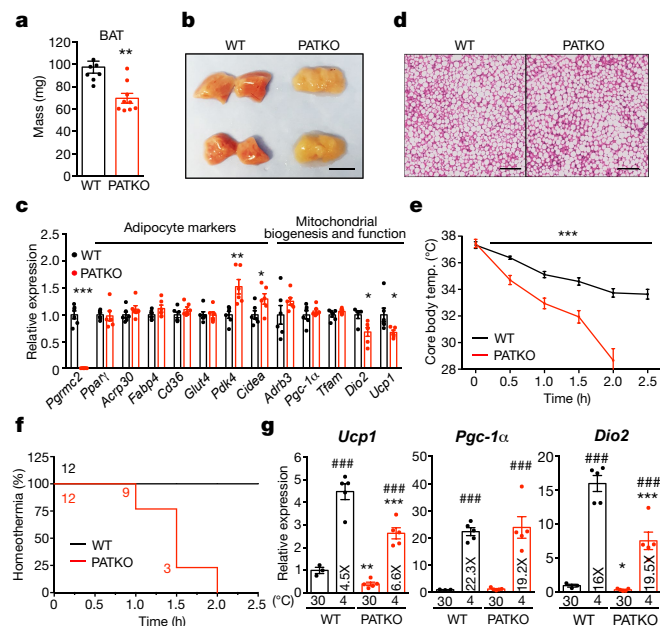
at 405 (ferric) and 400 (ferrous) nm.  $K_d$  is expressed as mean  $\pm$  s.d. **f**, Peroxidase activity in HEK293T cells co-transfected with labile haem reporters and scrambled or *Pgrmc2* siRNA, and exposed to succinylacetone (SA), haem-depleted FBS or both for 48 h. ER, endoplasmic reticulum. **g**, Endogenous PGRMC2 co-immunoprecipitates with endogenous PGRMC1 in primary brown adipocytes. **h**, Peroxidase activity in HEK293T cells co-transfected with labile haem reporters and scrambled, *Pgrmc1*, or *Pgrmc1* and *Pgrmc2* siRNA and treated as in **f**. The scrambled group is repeated from two (**b**, **g**) or three (**a**, **c**, **f**, **h**) independent samples. Representative results from two (**b**, **g**) or three (**a**, **c**, **f**, **h**) independent experiments. Data are presented as mean  $\pm$  s.d.; \* $P$  < 0.05 and \*\*\* $P$  < 0.001 versus scrambled basal; determined by two-way analysis of variance (ANOVA) with multiple comparisons and Tukey's post-test.

mutant (PGRMC2(3×M), Extended Data Fig. 1d, e), showed haem staining, indicating transfer of haem from PGRMC2 to apo-Rev-Erb $\alpha$  (Fig. 1d). Consistent with a role in serial trafficking<sup>16</sup>, PGRMC2 displayed medium-low affinity for haem (dissociation constant,  $K_d = 1.4 \times 10^{-6}$  M for ferric and  $5.3 \times 10^{-6}$  M for ferrous) (Fig. 1e). Total intracellular haem is the sum of haem covalently bound or nearly so as a cofactor, and labile—or signalling—haem, which is buffered by proteins and available for exchange and regulatory events<sup>4</sup>. To assess the ability of PGRMC2 to modulate subcellular labile haem levels, we transfected HEK293T cells with GFP–haemoprotein peroxidase fusion reporters targeted to mitochondria, endoplasmic reticulum, cytosol and nuclei<sup>17</sup> (Extended Data Fig. 1f). The activity of these reporters depends on the availability of labile haem in these compartments. Because intracellular labile haem may be derived from the medium or from endogenous synthesis, to examine the contribution of PGRMC2 to haem mobilization from either source, we used succinylacetone to block biosynthesis<sup>18</sup> and haem-depleted fetal bovine serum (FBS) to minimize exogenous haem uptake. Measurements in control cells showed that the mitochondrial and nuclear labile haem pools were derived entirely from endogenous synthesis, for the activity of these reporters was fully blunted in cells treated with succinylacetone (Fig. 1f). By contrast, both endogenous and exogenous haem contributed to the cytosolic and endoplasmic reticulum labile haem pools (Fig. 1f). PGRMC2 depletion resulted in decreased reporter activity in mitochondria, nuclei and—to a lesser extent—the endoplasmic reticulum, indicating reduced presence of labile haem (Fig. 1f, Extended Data Fig. 1g). We next considered how PGRMC2, which is localized in the endoplasmic reticulum and nuclear envelope, might acquire haem. Notably, PGRMC1 forms a complex with ferrochelatase (FECH) in the mitochondrial outer membrane that controls haem release<sup>19</sup>. PGRMC1 and PGRMC2 also interact<sup>20</sup>, and both are present in mitochondria-associated membranes<sup>21,22</sup>. We noted that in primary brown adipocytes PGRMC2 interacted with PGRMC1,

but not with glyceraldehyde-3-phosphate dehydrogenase, which was recently designated a cytosolic haem chaperone<sup>16</sup> (Fig. 1g). No interaction between PGRMC2 and PGRMC1 was detected when an antibody targeting the haem-binding domain was used, suggesting that PGRMC2 interacts with PGRMC1 at or near this region (Extended Data Fig. 1h). Depletion of PGRMC1 resulted in reduced labile haem in all subcellular compartments, probably reflecting its broader pattern of localization<sup>9</sup> (Fig. 1h). Dual knockdown of PGRMC1 and PGRMC2 had no added effect (Fig. 1h, Extended Data Fig. 1g), suggesting that PGRMC2 acts downstream of PGRMC1 to traffic endogenously synthesized haem. These findings suggest a model in which mitochondria-bound PGRMC1 transfers haem to endoplasmic-reticulum-bound PGRMC2, which delivers haem to proteins in the endoplasmic reticulum and nucleus, including haem-responsive transcription factors such as Rev-Erb $\alpha$ .

## PGRMC2 is required for thermogenesis

To evaluate the importance of PGRMC2-mediated haem mobilization in vivo, we focused on adipose tissue. PGRMC2 is enriched in adipose depots, particularly brown adipose tissue (BAT) (Extended Data Fig. 2a, b). We generated adipose-specific PGRMC2-null mice—which we designated PGRMC2 adipose tissue knockout (PATKO)—that lack PGRMC2 only in mature adipocytes. To avoid compensation mechanisms, unless noted all procedures were conducted at thermoneutrality. PATKO mice adapted to 30 °C showed no difference in body weight or white adipose tissue (WAT) mass (Extended Data Fig. 2c) but had reduced BAT weight (Fig. 2a) relative to their wild-type littermates. Notably, the appearance of PGRMC2-deficient BAT was markedly altered, with loss of its distinctive reddish colour (Fig. 2b). There was, however, no difference in expression of brown adipocyte markers (Fig. 2c) and histological comparison failed to reveal any difference (Fig. 2d). These findings led us to test the functionality of PGRMC2-deficient BAT. Reflecting the minor role



**Fig. 2 | PATKO mice are sensitive to cold.** **a**, **b**, Weight (**a**) and gross appearance (**b**) of BAT of chow-fed wild-type (WT) ( $n=8$ ) and PATKO ( $n=9$ ) mice maintained at 30 °C. Scale bar, 1 cm. **c**, Expression of thermogenic genes is decreased in PATKO BAT (wild type  $n=5$ ; PATKO  $n=6$ ). **d**, Haematoxylin and eosin (H&E) staining of BAT. Representative images from two independent experiments ( $n=5$ ). Scale bar, 100  $\mu$ m. **e**, PATKO mice are cold-intolerant ( $n=12$ ). Challenge started at Zeitgeber time (ZT)5. **f**, Survival curves at 4 °C (homeothermia is at 31 °C). **g**, PATKO BAT responds normally to adrenergic signalling (wild type,  $n=4$ ; PATKO,  $n=5$ ). *Pgc-1α* is also known as *Ppargc1a*. In **a–g**,  $n$  represents biologically independent samples. Data are mean  $\pm$  s.e.m. \* $P < 0.05$ , \*\* $P < 0.01$  and \*\*\* $P < 0.001$ ; PATKO versus wild type. \*\*\* $P < 0.001$ ; 30 °C versus 4 °C, determined by two-tailed Student's  $t$ -test (**a**, **c**) or two-way ANOVA with multiple comparisons and Bonferroni's post-test (**e**, **g**).

of BAT at thermoneutrality, PATKO mice were indistinguishable from wild-type mice in energy balance studies (Extended Data Fig. 2d). However, in contrast to wild-type mice, which activated thermogenesis and preserved body temperature when exposed to cold (4 °C), PATKO mice rapidly became hypothermic and perished if not rescued (Fig. 2e, f). This total impairment of adaptive thermogenesis was not a result of reduced sympathetic stimulation, as the transcriptional response to noradrenaline remained intact (Fig. 2g), despite a modest decrease in plasma noradrenaline (Extended Data Fig. 2e). Plasma glucose during challenge was similar to that of wild-type mice, and non-esterified fatty acids were minimally reduced (Extended Data Fig. 2e). To confirm that the thermogenic defect was independent of noradrenaline levels, we used the  $\beta_3$ -adrenergic receptor agonist CL316,243. Injection of CL316,243 elicited an immediate and sustained increase in oxygen consumption in wild-type mice; this response was significantly blunted in PATKO mice (Extended Data Fig. 2f). Further, consistent with our model of mitochondrial haem mobilization, adipose-specific PGRMC1 and PGRMC2 double-knockout mice were also cold-sensitive, perhaps more so than PATKO mice (Extended Data Fig. 2g). These findings stress the importance of the PGRMC1–PGRMC2 haem-trafficking pathway for adaptive thermogenesis.

### Loss of PGRMC2 causes mitochondrial dysfunction

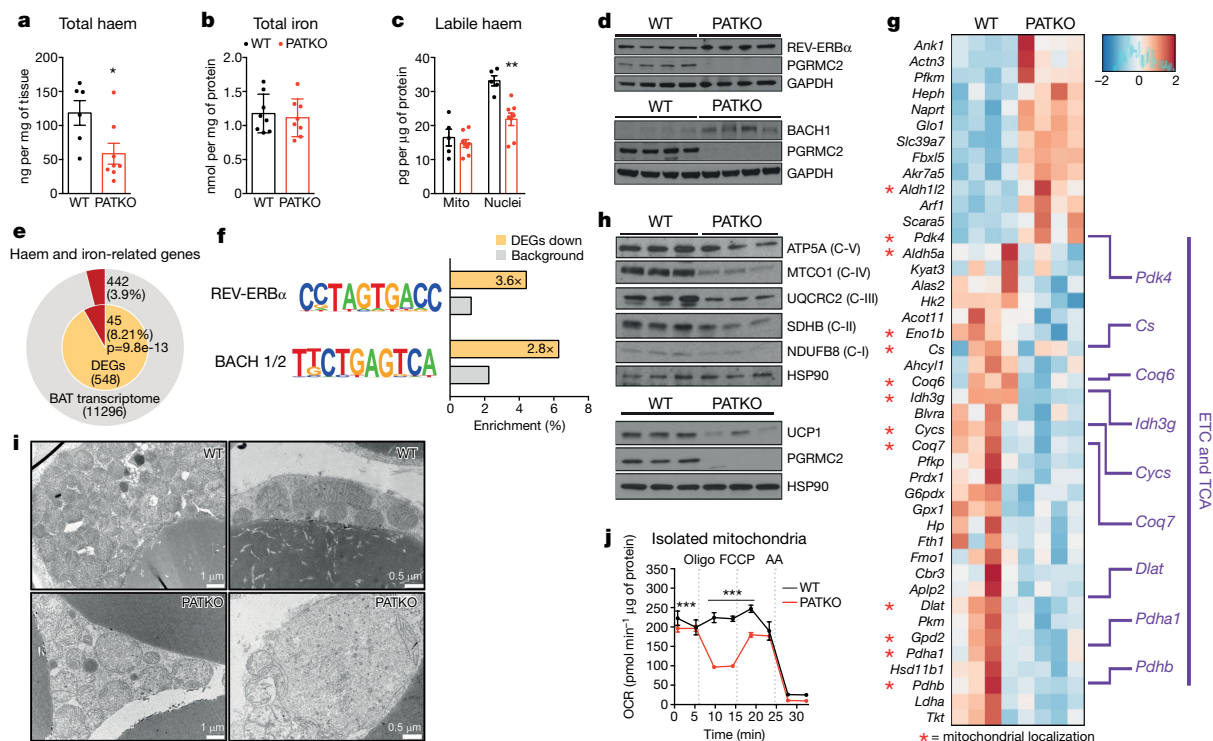
To determine the basis of the defects of PGRMC2-null BAT, we measured total haem content and found it considerably reduced (about 60%) (Fig. 3a). To probe the origins of this difference, we quantified haem precursors and found reduced levels of succinyl-CoA and glycine, the substrates of 5'-aminolevulinic acid synthase 1 (ALAS1), the rate-limiting enzyme of haem biosynthesis (Extended Data Fig. 3a). Accordingly,

levels of 5-aminolevulinic acid, the product of ALAS1, tended to decrease (Extended Data Fig. 3a). We also noted decreased expression of *Alas1* and *Alas2* (Extended Data Fig. 3b), indicating that defects in biosynthesis contribute to decreased total haem in PATKO BAT. Iron content was the same as in wild-type BAT, indicating that the reduced haem levels were not caused by iron deficiency (Fig. 3b) and suggesting that tissue haem uptake was unaffected. Of note, labile haem levels were significantly decreased in nuclei purified from PATKO (Fig. 3c) and PGRMC1 and PGRMC2 double-knockout BAT, which had a similar discoloured appearance to PATKO BAT (Extended Data Fig. 3c). In the nucleus, haem regulates the activity of several transcription factors that, upon binding haem, are ultimately degraded. These include Rev-Erba and the transcriptional repressor BACH1<sup>23,24</sup>. Levels of Rev-Erba and BACH1 proteins were higher in PATKO BAT (Fig. 3d), indicating that reduced nuclear labile haem resulted in stabilization of these factors. Accordingly, expression of *Bmal1* (also known as *Arntl*) and *Fth1*, targets of Rev-Erba and BACH1, respectively, was reduced (Extended Data Fig. 3d). The circadian pattern of Rev-Erba mRNA expression<sup>25</sup> was not altered in PATKO mice (Extended Data Fig. 3e), suggesting that the increased Rev-Erba protein levels in PATKO BAT are probably the result of reduced degradation. RNA-sequencing (RNA-seq) analysis showed that among differentially expressed genes (DEGs) between wild-type and PATKO BAT (adjusted  $P < 0.05$ ; 312 DEGs upregulated and 236 DEGs downregulated) (Supplementary Table 1), haem and iron homeostasis genes were enriched (45 genes, 8.2% of DEGs versus 3.9% in the BAT transcriptome;  $P < 10^{-13}$ ) (Fig. 3e). Enhancer analysis of downregulated DEGs in PATKO BAT revealed an enrichment ( $P < 10^{-7}$ ) of Rev-Erba and BACH1 and BACH2 motifs (Fig. 3f, Supplementary Table 2), consistent with altered regulation of haem-sensitive transcription. The majority of haem and iron-linked DEGs were present in the three most-downregulated pathways, which relate to metabolic processes and energy generation and contain many mitochondrial proteins. Expression of electron transport chain and tricarboxylic acid cycle genes (Extended Data Fig. 3f, g) was broadly decreased in PATKO BAT, and levels of all electron transport chain proteins analysed were notably lower (Fig. 3g, h). Further, PATKO BAT had substantially reduced levels of uncoupling protein 1 (UCP1) (Fig. 3h), a finding consistent with greater stability of Rev-Erba, which directly represses *Ucp1*<sup>26</sup>. Beyond its role in uncoupling mitochondrial electron transport, UCP1 regulates mitochondrial integrity<sup>27</sup>. PGRMC2-null brown adipocytes have large, swollen mitochondria with few, disorganized cristae (Fig. 3i), indicating mitochondrial dysfunction. Indeed, mitochondria isolated from PATKO BAT had reduced basal and markedly reduced uncoupled respiration (Fig. 3j). These findings demonstrate that in the absence of PGRMC2 there is a lower level of labile haem in the nucleus, leading to changes in the haem-responsive transcriptome that cause mitochondrial dysfunction.

### Endogenous haem controls mitochondrial function

Primary brown PATKO adipocytes recapitulated these defects: they exhibited severely reduced respiratory capacity, a markedly blunted response to adrenergic stimuli without alterations in the transcriptional response to noradrenaline, and decreased levels of UCP1 and electron transport chain proteins (Extended Data Fig. 4a–j). Similar, and perhaps greater, defects were noted in adipocytes deficient in both PGRMC1 and PGRMC2 (Extended Data Fig. 4k). The introduction of human PGRMC2 into mouse PGRMC2-null brown adipocytes restored mitochondrial bioenergetics and UCP1 levels, whereas expression of a PGRMC2 haem-binding mutant did not, indicating that these defects are related to the ability of PGRMC2 to mobilize haem (Extended Data Fig. 4l–o). Notably, mirroring the effect of PGRMC2 deletion, inhibition of haem synthesis was sufficient to impair mitochondrial function and deplete UCP1 in wild-type cells (Extended Data Fig. 5a–d). Neither depletion of exogenous haem nor addition of haemin affected mitochondrial





**Fig. 3 | PGRMC2 regulates haem-sensitive transcription and mitochondrial function in BAT.** **a**, **b**, Total haem (**a**; wild type,  $n = 6$ ; PATKO,  $n = 8$ ) and iron (**b**;  $n = 8$ ) levels in BAT. **c**, Labile haem in mitochondrial and nuclear fractions of BAT (wild type,  $n = 5$ ; PATKO,  $n = 8$ ). **d**, Rev-Erbα and BACH1 levels in BAT. **e**, Genes related to haem and iron metabolism (red portions) are enriched in DEGs. **f**, Rev-Erbα- and BACH1- and BACH2-binding motifs are enriched in genes downregulated in PATKO BAT. **g**, Heat map of haem- and iron-related genes shows a global decrease of electron transport chain (ETC) and tricarboxylic acid

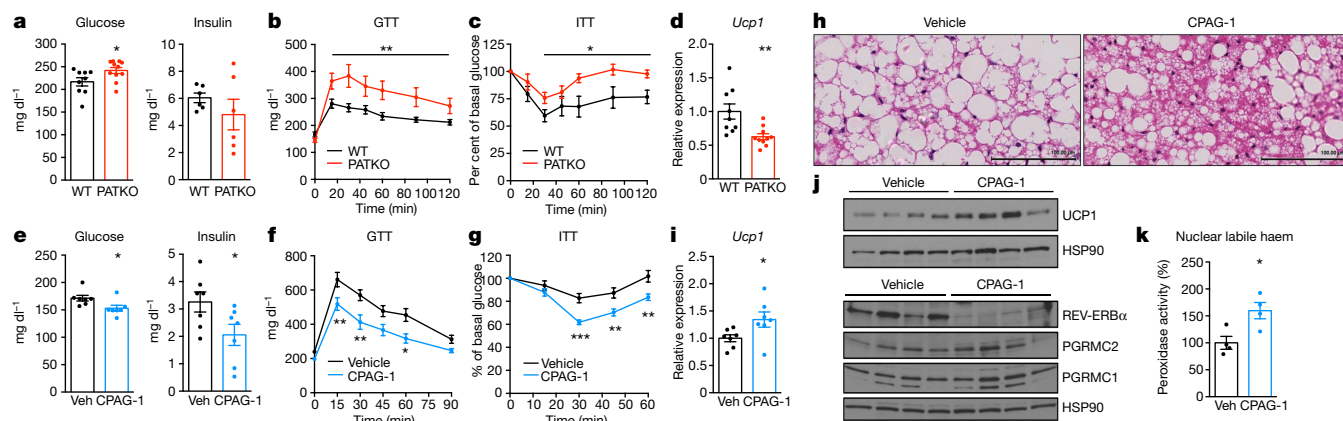
respiration in wild-type or PATKO cells (Extended Data Fig. 5a–e). These observations show that PGRMC2-dependent mobilization of endogenous haem regulates mitochondrial function in brown adipocytes. Lastly, we found that both Rev-Erbα and BACH1 proteins were more abundant in PATKO cells (Extended Data Fig. 5f). Dual knockdown of these factors restored basal respiration in PGRMC2-null adipocytes (Extended Data Fig. 5g, h), indicating that they are

key mediators of the transcriptional response to haem and its effect on mitochondrial function. **h**, UCP1 and oxidative phosphorylation (OXPHOS) proteins are reduced in PATKO BAT. **i**, Electron microscopy shows altered mitochondrial morphology in PATKO BAT. Representative images from four biologically independent samples. **j**, Oxygen consumption rate (OCR) of mitochondria isolated from BAT ( $n = 6$ ). In **a–j**,  $n$  represents biologically independent samples. Representative results from two (**a–c**, **j**) or three (**d**, **h**) independent experiments. Data are mean  $\pm$  s.e.m. \* $P < 0.05$ , \*\* $P < 0.01$  and \*\*\* $P < 0.001$  versus wild type; by two-tailed Student's  $t$ -test.

cycle (TCA) gene expression. **h**, UCP1 and oxidative phosphorylation (OXPHOS) proteins are reduced in PATKO BAT. **i**, Electron microscopy shows altered mitochondrial morphology in PATKO BAT. Representative images from four biologically independent samples. **j**, Oxygen consumption rate (OCR) of mitochondria isolated from BAT ( $n = 6$ ). In **a–j**,  $n$  represents biologically independent samples. Representative results from two (**a–c**, **j**) or three (**d**, **h**) independent experiments. Data are mean  $\pm$  s.e.m. \* $P < 0.05$ , \*\* $P < 0.01$  and \*\*\* $P < 0.001$  versus wild type; by two-tailed Student's  $t$ -test.

## Adipose PGRMC2 regulates systemic metabolism

We next gauged the importance of adipose PGRMC2 for glucose homeostasis. PATKO mice housed at room temperature and fed a high-fat diet



**Fig. 4 | PGRMC2 controls systemic glucose homeostasis.** **a**, Blood glucose (wild type,  $n = 9$ ; PATKO,  $n = 10$ ) and insulin ( $n = 6$ ) in wild-type and PATKO mice on HFD. **b**, **c**, Glucose tolerance test (GTT) (**b**) and insulin tolerance test (ITT) (**c**) after 10 (GTT) and 12 (ITT) weeks of HFD (GTT: wild type,  $n = 8$ , PATKO,  $n = 11$ ; ITT: wild type,  $n = 6$ , PATKO,  $n = 9$ ). **d**, *Ucp1* mRNA in BAT of HFD-fed wild-type ( $n = 9$ ) and PATKO ( $n = 10$ ) mice. **e**, Glucose and insulin levels in DIO mice treated with vehicle or CPAG-1 for 30 days ( $n = 7$ ). **f**, **g**, GTT (**f**) and ITT (**g**) in DIO mice after 14 (GTT) and 20 (ITT) days of treatment ( $n = 7$ ). **h**, H&E staining of BAT. Representative images from four biologically independent samples. **i**, *Ucp1* mRNA levels in BAT of treated DIO mice ( $n = 7$ ). **j**, UCP1 and Rev-Erbα levels in BAT of treated DIO mice. **k**, Nuclear labile haem levels in BAT of DIO mice treated with CPAG-1 for four days ( $n = 4$ ). In **a–k**,  $n$  represents biologically independent samples, representative results from two independent experiments. Data are mean  $\pm$  s.e.m. \* $P < 0.05$ , \*\* $P < 0.01$  and \*\*\* $P < 0.001$  versus wild type or vehicle; two-tailed Student's  $t$ -test (**a**, **d**, **e**, **i**, **k**) or two-way ANOVA with multiple comparisons and a Bonferroni's post-test (**b**, **c**, **f**, **g**).

Representative images from four biologically independent samples. **i**, *Ucp1* mRNA levels in BAT of treated DIO mice ( $n = 7$ ). **j**, UCP1 and Rev-Erbα levels in BAT of treated DIO mice. **k**, Nuclear labile haem levels in BAT of DIO mice treated with CPAG-1 for four days ( $n = 4$ ). In **a–k**,  $n$  represents biologically independent samples, representative results from two independent experiments. Data are mean  $\pm$  s.e.m. \* $P < 0.05$ , \*\* $P < 0.01$  and \*\*\* $P < 0.001$  versus wild type or vehicle; two-tailed Student's  $t$ -test (**a**, **d**, **e**, **i**, **k**) or two-way ANOVA with multiple comparisons and a Bonferroni's post-test (**b**, **c**, **f**, **g**).

(HFD) showed no differences in body weight or composition, except for decreased BAT mass (Extended Data Fig. 6a, b). However, they had higher fasting glycaemia (Fig. 4a) and decreased glucose tolerance and insulin sensitivity (Fig. 4b, c). They also exhibited hyperlipidaemia and exacerbated liver steatosis (about 70% more triglycerides) (Extended Data Fig. 6c–e), factors that probably increased insulin resistance. The BAT of HFD-fed PATKO mice showed no histological abnormalities (Extended Data Fig. 7a) but had substantially reduced *Ucp1* expression (approximately 40% less) (Fig. 4d). Expression of *Bmal1* and *Fth1* was also decreased (Extended Data Fig. 7b). Analysis of WAT depots did not show extensive differences in adipocyte size, immune cell infiltration or gene expression in inguinal or epididymal WAT (Extended Data Fig. 7c–e). Notably, *Bmal1* expression was reduced in PATKO inguinal WAT (Extended Data Fig. 7e). We propose that hastened metabolic deterioration in HFD-fed PATKO mice probably reflects the aggregate of defects in BAT and WAT.

### PGRMC2 activation mitigates metabolic disease

The deleterious effects on metabolism of adipose PGRMC2 deletion suggest that activation of PGRMC2 function might reverse features of metabolic syndrome. Thus, we treated diet-induced-obese (DIO) mice at room temperature with a small-molecule PGRMC2 activator (compound 27 in ref. <sup>5</sup>; hereafter referred to as CPAG-1). CPAG-1 treatment had no effect on weight or food intake (Extended Data Fig. 8a), but treated mice had reduced fasting glycaemia and insulin levels (Fig. 4e) and improved glucose tolerance and insulin sensitivity (Fig. 4f, g). BAT histology showed decreased lipid content and an increase in multilocular adipocytes (Fig. 4h), features indicative of improved function. Expression of *Ucp1* and *Bmal1* was also upregulated (Fig. 4i, Extended Data Fig. 8b), changes suggestive of reduced levels of Rev-Erba. Indeed, Rev-Erba protein was decreased and UCP1 protein was increased in BAT of CPAG-1-treated mice (Fig. 4j). Labile haem in the nucleus of brown adipocytes from CPAG-1-treated mice was significantly increased within four days of treatment (Fig. 4k), suggesting that decreased Rev-Erba protein was probably the result of haem-induced degradation. No histological differences were found in inguinal WAT (Extended Data Fig. 8c), but expression of *Ucp1* and *Pgc-1α* was increased (Extended Data Fig. 8d). Histology revealed a marked improvement in epididymal WAT, with fibrosis and inflammation noticeably decreased (Extended Data Fig. 8e, f). The liver of CPAG-1-treated mice appeared slightly less steatotic and expression of gluconeogenic genes and *Tnfa* was reduced (Extended Data Fig. 8g, h). CPAG-1 treatment also increased hepatic nuclear labile haem levels (Extended Data Fig. 8i). Given that CPAG-1 interacts very weakly with PGRMC1<sup>5</sup> (Extended Data Fig. 9), we suggest it may act primarily through PGRMC2 to increase haem flux to the nucleus.

### Discussion

In this study we have described a role for PGRMC2 in transport of mitochondrial haem. In the absence of PGRMC2, less labile haem reaches the nucleus, resulting in alterations in haem-sensitive transcription that cause mitochondrial dysfunction in brown adipocytes (Extended Data Fig. 10). These defects compromise not only the primary function of BAT (preservation of normal body temperature), but also its contribution to systemic glucose homeostasis. Given its high expression across white fat depots, further studies will be needed to determine whether PGRMC2 performs a similar role in WAT. Nevertheless, our findings provide a view of how haem dynamics in adipocytes can affect physiology. Haem levels and expression of biosynthetic enzymes are reduced in visceral fat of obese humans<sup>28</sup>, stressing the link between adipocyte haem homeostasis and metabolic disease. Because PGRMC2 is restricted in its tissue distribution, additional haem chaperones probably remain to be discovered. Finally, we have shown that pharmacological activation of PGRMC2 may be of use in treating metabolic disease. Given the interest in identifying signalling pathways that enhance adipocyte function and

correct obesity-linked adipose tissue defects<sup>29</sup>, our findings suggest that modulation of intracellular haem dynamics could be a potentially innovative therapeutic strategy.

### Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1774-2>.

- Severance, S. & Hamza, I. Trafficking of heme and porphyrins in metazoa. *Chem. Rev.* **109**, 4596–4616 (2009).
- Mense, S. M. & Zhang, L. Heme: a versatile signaling molecule controlling the activities of diverse regulators ranging from transcription factors to MAP kinases. *Cell Res.* **16**, 681–692 (2006).
- Donegan, R. K., Moore, C. M., Hanna, D. A. & Reddi, A. R. Handling heme: the mechanisms underlying the movement of heme within and between cells. *Free Radic. Biol. Med.* **133**, 88–100 (2019).
- Reddi, A. R. & Hamza, I. Heme mobilization in animals: a metallolipid's journey. *Acc. Chem. Res.* **49**, 1104–1110 (2016).
- Parker, C. G. et al. Ligand and target discovery by fragment-based screening in human cells. *Cell* **168**, 527–541 (2017).
- Gerdes, D., Wehling, M., Leube, B. & Falkenstein, E. Cloning and tissue expression of two putative steroid membrane receptors. *Biol. Chem.* **379**, 907–911 (1998).
- Wendler, A. & Wehling, M. PGRMC2, a yet uncharacterized protein with potential as tumor suppressor, migration inhibitor, and regulator of cytochrome P450 enzyme activity. *Steroids* **78**, 555–558 (2013).
- Jühlen, R., Landgraf, D., Huebner, A. & Koehler, K. Identification of a novel putative interaction partner of the nucleoporin ALADIN. *Biol. Open* **5**, 1697–1705 (2016).
- Kimura, I. et al. Functions of MAPR (membrane-associated progesterone receptor) family members as heme/steroid-binding proteins. *Curr. Protein Pept. Sci.* **13**, 687–696 (2012).
- Chen, J. J. & London, I. M. Hemin enhances the differentiation of mouse 3T3 cells to adipocytes. *Cell* **26**, 117–122 (1981).
- Chawla, A. & Lazar, M. A. Induction of Rev-Erba, an orphan receptor encoded on the opposite strand of the α-thyroid hormone receptor gene, during adipocyte differentiation. *J. Biol. Chem.* **268**, 16265–16269 (1993).
- Kojet, D. J. & Burris, T. P. A role for Rev-Erba ligands in regulation of adipogenesis. *Curr. Pharm. Des.* **17**, 320–324 (2011).
- Wang, J. & Lazar, M. A. Bifunctional role of Rev-Erba in adipocyte differentiation. *Mol. Cell. Biol.* **28**, 2213–2220 (2008).
- Raghuram, S. et al. Identification of heme as the ligand for the orphan nuclear receptors REV-ERBA and REV-ERBβ. *Nat. Struct. Mol. Biol.* **14**, 1207–1213 (2007).
- Yin, L. et al. Rev-Erba, a heme sensor that coordinates metabolic and circadian pathways. *Science* **318**, 1786–1789 (2007).
- Sweeny, E. A. et al. Glyceraldehyde-3-phosphate dehydrogenase is a chaperone that allocates labile heme in cells. *J. Biol. Chem.* **293**, 14557–14568 (2018).
- Yuan, X. et al. Regulation of intracellular heme trafficking revealed by subcellular reporters. *Proc. Natl Acad. Sci. USA* **113**, E5144–E5152 (2016).
- Ebert, P. S., Hess, R. A., Frykholm, B. C. & Tschudy, D. P. Succinylacetone, a potent inhibitor of heme biosynthesis: effect on cell growth, heme content and δ-aminolevulinic acid dehydratase activity of malignant murine erythroleukemia cells. *Biochem. Biophys. Res. Commun.* **88**, 1382–1390 (1979).
- Piel, R. B., III et al. A novel role for progesterone receptor membrane component 1 (PGRMC1): a partner and regulator of ferrochelatase. *Biochemistry* **55**, 5204–5217 (2016).
- Peluso, J. J., Griffin, D., Liu, X. & Horne, M. Progesterone receptor membrane component-1 (PGRMC1) and PGRMC-2 interact to suppress entry into the cell cycle in spontaneously immortalized rat granulosa cells. *Biol. Reprod.* **91**, 104 (2014).
- Hung, V. et al. Proteomic mapping of cytosol-facing outer mitochondrial and ER membranes in living human cells by proximity biotinylation. *eLife* **6**, e24463 (2017).
- Medlock, A. E. et al. Identification of the mitochondrial heme metabolism complex. *PLoS ONE* **10**, e0135896 (2015).
- Carter, E. L., Gupta, N. & Ragsdale, S. W. High affinity heme binding to a heme regulatory motif on the nuclear receptor Rev-Erba leads to its degradation and indirectly regulates its interaction with nuclear receptor corepressor. *J. Biol. Chem.* **291**, 2196–2222 (2016).
- Zenke-Kawasaki, Y. et al. Heme induces ubiquitination and degradation of the transcription factor Bach1. *Mol. Cell. Biol.* **27**, 6962–6971 (2007).
- Everett, L. J. & Lazar, M. A. Nuclear receptor Rev-Erba: up, down, and all around. *Trends Endocrinol. Metab.* **25**, 586–592 (2014).
- Gerhart-Hines, Z. et al. The nuclear receptor Rev-Erba controls circadian thermogenic plasticity. *Nature* **503**, 410–413 (2013).
- Kazak, L. et al. UCP1 deficiency causes brown fat respiratory chain depletion and sensitizes mitochondria to calcium overload-induced dysfunction. *Proc. Natl Acad. Sci. USA* **114**, 7981–7986 (2017).
- Moreno-Navarrete, J. M. et al. Heme biosynthetic pathway is functionally linked to adipogenesis via mitochondrial respiratory activity. *Obesity* **25**, 1723–1733 (2017).
- Kajimura, S. & Saito, M. A new era in brown adipose tissue biology: molecular control of brown fat development and energy homeostasis. *Annu. Rev. Physiol.* **76**, 225–249 (2014).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019



## Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

### Reagents

Haemin, protoporphyrin IX, oligomycin A, carbonyl cyanide 4-(trifluoromethoxy) phenylhydrazone (FCCP), rotenone, antimycin A, 3-isobutyl-1-methylxanthine (IBMX), BSA, mannitol, noradrenaline, isoproterenol, 8-Br-cAMP and succinylacetone were purchased from Sigma-Aldrich. CL 316,243 was obtained from Cayman Chemical. Forskolin was obtained from Chem Impex International. Insulin (Novolin) was purchased from Novo-Nordisk. Complete EDTA-free protease inhibitor cocktail was obtained from Roche. DMEM and other Gibco-branded cell culture products were purchased from Thermo Fisher. Haem-depleted FBS was prepared by treating FBS with 20 mM ascorbic acid for 16 h, followed by 24 h dialysis against PBS. Haem depletion was verified by measuring optical absorbance at 405 nm. CPAG-1 was synthesized as previously reported<sup>5</sup>. ON-TARGET siRNA SMARTpools against human *PGRMC2* (L-010639-00-0005), *PGRMC1* (L-010642-00-0005), and mouse *Nr1d1* (L-051721-00-0005), and *BACH1* (L-042956-01-0005), as well as a Non-targeting Pool (D-001810-10-05) were purchased from Dharmacon. HEK293T cells were obtained from ATCC (CRL-3216) having undergone short-tandem repeat verification. Cells were routinely tested for mycoplasma and were never positive.

### Protein production

Full-length PGRMC2 in a bacterial expression vector (GenBank Accession number NM\_027558; Genecopoeia Ex-Mm25103-B01) was transformed into chemically competent BL21(DE3) cells (Thermo Fisher) and grown at 37 °C to an OD<sub>600</sub> of 0.8. Cells were induced with 1 mM IPTG and grown for 12 h at 30 °C. Cells were collected and stirred at room temperature in 50 mM Tris-HCl, 150 mM NaCl pH 8.5 containing 1% Triton X-100, 100 µg/ml lysozyme, 100 µg/ml DNase I, 10 mM MgCl<sub>2</sub>, and 10 mM CaCl<sub>2</sub>, and 1× Complete EDTA-free protease inhibitor cocktail (Roche) for 1 h. After sonication, the lysate was centrifuged at 6,000g for 30 min and the supernatant purified using nickel affinity chromatography. After elution, the protein was dialysed into 50 mM Tris-HCl, 150 mM NaCl, pH 7.4, and purified by HiLoad 16/600 Superdex 75 size exclusion chromatography (GE Healthcare). A mouse PGRMC2 haem-binding mutant was created by mutating 3 amino acids (Y131F, K187A and Y188F) using a Quikchange II XL Site-Directed Mutagenesis Kit (Agilent), verified by DNA sequencing, and expressed and purified as described above. To generate the PGRMC2 cytochrome *b5* haem-binding domain, residues 102–209 of human PGRMC2 were codon-optimized, synthesized (Integrated DNA Technologies) and inserted into the pET21a vector. The plasmid construct was transformed into BL21(DE3) cells (Thermo Fisher). Cells were grown at 37 °C to an OD<sub>600</sub> of 1.0 and induced with 1 mM IPTG for 5 h, collected and resuspended in 50 mM Tris-HCl, 1 mM EDTA, 0.01% Na<sub>2</sub>S<sub>2</sub>O<sub>3</sub>, 1 mM DTT, 25% sucrose, and lysed in 50 mM Tris-HCl, 1 mM EDTA, 0.01% Na<sub>2</sub>S<sub>2</sub>O<sub>3</sub>, 1 mM DTT, 200 mM sodium chloride, 1% sodium deoxycholate and 1% Triton X-100. To isolate inclusion bodies, lysed cells were centrifuged at 6,000g for 20 min and washed extensively with 50 mM Tris-HCl, 1 mM EDTA, 0.01% Na<sub>2</sub>S<sub>2</sub>O<sub>3</sub>, 1 mM DTT, 25% sucrose, 100 mM sodium chloride and 0.5% Triton X-100. Inclusion bodies were subjected to a final wash in the same buffer without Triton X-100. To denature inclusion bodies, about 200 mg of inclusion bodies was resuspended in 100 mM Tris-HCl, 6 M guanidinium chloride and 20 mM β-mercaptoethanol for 1 h at room temperature (RT). Denatured inclusion bodies were refolded overnight at 4 °C in an oxidative refolding buffer containing 400 mM L-arginine, 100 mM Tris-HCl, 5 mM reduced glutathione, 0.5 mM oxidized glutathione, 10 mM EDTA and 200 mM phenylmethylsulphonyl fluoride. Refolded protein was concentrated and purified using HiLoad 16/600 Superdex 75 size

exclusion chromatography (GE Healthcare). Purity of PGRMC2 proteins was confirmed using SDS-PAGE. Soret and α, β absorption spectra were measured on a SpectraMAX 250 reader (Molecular Devices) at room temperature. Purified PGRMC2 protein was incubated for 15 min with 10 mM dithionite to reduce the haem group. Human REV-ERBα LBD (residues 281–614) with an N-terminal hexahistidine tag and a tobacco etch virus (TEV) protease cleavage site was inserted into a pET46 vector and expressed in *E. coli* BL21(DE3) cells. Cells were grown in at 37 °C overnight and induced in autoinduction medium at 37 °C for 5 h, 30 °C for 1 h, and 22 °C for 16 h. Cells were collected and pellets stored at –80 °C. Pellets were thawed on ice and resuspended in lysis buffer without imidazole (40 mM NaHPO<sub>4</sub>, pH 7.4, 500 mM NaCl, 10% glycerol, 2.5 mM DTT and 0.1% Tween-20) at 40 ml buffer per 5 g pellet. The cell slurry was sonicated on ice in 15 s on/30 s off intervals (75% amplitude) for 5 min total. Lysed cells were centrifuged at 14,000 rpm for 30 min at 4 °C. The supernatant was filtered through a 0.4-µm PES membrane Nalgene Rapid-Flow bottle-top filter and affinity purified using 2 × 5 ml HisTrap IMAC columns (GE Healthcare) affixed to an Äkta Start. After loading, columns were washed with 100 ml wash buffer (40 mM NaHPO<sub>4</sub>, pH 7.4, 500 mM NaCl, 10% glycerol, 15 mM imidazole and 1 mM DTT). The protein was eluted using a 10 column-volume elution gradient with elution buffer (40 mM NaHPO<sub>4</sub>, pH 7.4, 500 mM NaCl, 10% glycerol, 500 mM imidazole and 1 mM DTT). The protein was eluted after >50% elution buffer then pooled and dialysed in 10-kDa MWCO SnakeSkin dialysis tubing (Thermo Fisher) for 24 h at 4 °C to remove imidazole and bound haem in 2 l dialysis buffer (40 mM NaHPO<sub>4</sub>, pH 7.4, 500 mM NaCl, 10% glycerol, 10 mM DTT, 0.1% Tween-20 and 0.5 mM EDTA). After dialysis, the protein was concentrated using a 30 kDa MWCO Amicon Ultra centrifugal concentrator (EMD Millipore). The protein was further purified by size exclusion chromatography (Äkta Pure) using a Superdex 75 10/300 GL column in gel filtration buffer (20 mM NaHPO<sub>4</sub>, pH 7.4, 50 mM NaCl, 50 mM L-arginine, 50 mM L-glutamate and 0.5 mM EDTA). The protein was pooled and confirmed to be >90% pure by LC-MS and SDS-PAGE. The fraction of final purified 6×His-REV-ERBα LBD bound to haem was assessed using the extinction coefficient for the haem Soret peak ( $\epsilon_{415}$ ) of 101.85 = 1 mM, and 6×His-REV-ERBα LBD was confirmed to be >95% haem-free.

### Haem titration assay

The affinity of PGRMC2 cytochrome *b5* haem-binding domain for ferric and ferrous haem was measured by spectroscopy of the UV-visible spectrum in the Soret region using a SpectraMAX 250 reader. Sequential aliquots of haemin in DMSO were added to the sample well containing 10 µM apo-PGRMC2 and the reference well to obtain a 2-µM increment of haemin concentration per addition. Spectra were recorded 3 min after each addition of haemin. The difference in absorbance at 420 nm was plotted in relation to haemin concentration, and dissociation constants ( $K_d$ ) calculated with GraphPad Prism 6 using a quadratic binding equation.

### Haem transfer assay

Twenty-five microlitres of 200 nM apo-HRP (Calzyme Laboratories) was incubated with 25 µl of 5 µM purified PGRMC2 protein. After 5 min at room temperature, 150 µl of BioFX TMB One Component HRP microwell Substrate (Surmodics) was added to wells and absorbance at 405 nm measured immediately for 15 min. As a positive control, apo-HRP was incubated with 0.3 nM haemin for 5 min and absorbance measured as described.

### Native PAGE and in-gel haem staining

Haem transfer was assessed by mixing 10 µg of wild-type or mouse PGRMC2 haem-binding mutant (3×M) with 10 µg of apo-REV-ERBα protein and incubating for 30 min at 37 °C. After incubation, 2× Native Tris-Glycine sample buffer (Life Technologies) was added and samples separated by electrophoresis using Novex Tris-Glycine 4–20% gels and

## Article

Tris-Glycine Native Running Buffer (Life Technologies) for 6 h. The gel was washed for 10 min with water and haem staining was performed using the BioFX TMB One Component HRP Microwell Substrate (Surmodics). After imaging the haem stain, the gel was washed overnight with water and counterstained with Coomassie for protein detection.

### Mass spectrometry

To detect haem in purified PGRMC2 protein, 5 µl of 20 mg/ml PGRMC2 were extracted with 1 ml of Folch solution (2:1 chloroform:methanol) and washed with 200 µl of water. The extraction solution was then vortexed and centrifuged at 1,000g, 4 °C for 10 min and the lower phase extracted and dried down. Before LC-MS analysis, the sample was reconstituted in methanol. A haemin standard solution was prepared at 10 µM. LC-MS analysis was performed on an I-class UPLC system coupled with a Synapt G2-Si mass spectrometer via an electrospray ionization (ESI) source from Waters. The positive-mode (+) ESI conditions were as follows: capillary, +3.00 kV; sampling cone, 40 V; source temperature, 100 °C; desolvation temperature, 250 °C; desolvation gas flow, 600 l/h; and cone gas flow, 50 l/h, respectively. Leucine-enkephalin (*m/z* 556.2771) was used for lock mass correction. Liquid chromatography was performed with A = 40:60 water:acetonitrile + 1mM ammonium formate, B = 90:10 2-propanol:acetonitrile. A Waters ACQUITY UPLC BEH C18 column (1.7 µm, 2.1 mm × 100 mm) was used at a flow rate of 250 µl/min. Initially, the mobile phase composition consisted of 32% B and held for 1 min after injection and its composition was increased over the length of the gradient (15 min, B = 97%) in short increments adapted from a previous study<sup>30</sup>. The injection volume was 2 µl. For haem quantification in tissue, BAT was isolated from wild-type and PATKO mice housed at 30 °C after 10 min of perfusion with cold PBS. Ten to twenty-five milligrams of frozen tissue was homogenized in 300 µl of 1% formic acid in dH<sub>2</sub>O and an internal standard added. Haem was extracted in Folch solution (2:1 chloroform:methanol). After centrifugation at 4,000g for 10 min at 4 °C, haem was re-extracted from the organic phase with 1 volume of 1.4 N NaOH. Samples were centrifuged at 4,000g for 10 min at 4 °C and the aqueous phase collected for mass spectrometry analysis. Haemin was quantified on an Agilent 6495 triple quadrupole with a jet stream source coupled to an Agilent 1290 UPLC. As internal standard, deuteroporphyrin (Frontier Scientific) was used and the monitored transitions were *m/z* 616.1 → 557.1 (quantitative), *m/z* 616.1 → 498.2 (qualitative) for haemin, and *m/z* 564.0 → 505.0 for deuteroporphyrin. Jet stream was set at gas temperature 200 °C, gas flow 12 l/min, nebulizer pressure 30 psi, sheath gas temperature 325 °C, sheath gas flow 10 l/min, cap V = 400 V, nozzle V = 2,000 V. Liquid chromatography was performed with A = 90:10 water:methanol + 0.1% ammonium hydroxide and + 10 mM ammonium formate, B = 65:30:10 2-propanol:methanol:water + 0.1% ammonium hydroxide and + 10 mM ammonium formate. All solvents were LC-MS grade. An Agilent extend-C18 column (1.8 µm, 2.1 × 50 mm) was used at a flow rate of 0.2 ml/min. Initially, the mobile phase consisted of 5% B and, after injection, its composition increased linearly to 95% B in 6 min and held at 95% for 3 min. The injection volume was 5 µl. Haem content was normalized per milligram of tissue. Quantitative analysis of glycine, aminolevulinic acid and succinyl-CoA was performed using a QQQ mass spectrometer operated in positive-ion mode (Xevo TQ-XS from Waters). In brief, 10 mg of frozen BAT was homogenized with ice cold 80% methanol and glass beads and incubated on ice for additional 10 min. The tissue lysate was centrifuged at 18,000g for 10 min at 4 °C and split into two aliquots followed by drying down in a vacuum concentrator and stored at -80 °C before LC-MS/MS analysis. For glycine and aminolevulinic acid analysis, an aliquot was reconstituted in 1:1 acetonitrile:water and injected into a Waters ACQUITY UPLC BEH Amide column (1.7 µm, 2.1 mm × 100 mm) at a flow rate of 400 µl/min. The mobile phases consisted of A = water + 0.1% formic acid and B = acetonitrile + 0.1% formic acid. Initially, the mobile phase composition consisted of 95% B and held for 1 min after injection and its composition was decreased to 65%

over 6 min and then to 40% over 3 min and held for an additional 1 min. The following quantifier and qualifier transitions (collision energy in eV) were used for each metabolite: glycine: 76.0 → 30.3 (6 eV) and 48.2 (4 eV); <sup>13</sup>C-glycine: 78.0 → 31.0 (6 eV) and 49.0 (4 eV); aminolevulinic acid: 132.2 → 55.1 (18 eV), 68.3 (18 eV), 86.0 (10 eV), 114.0 (6 eV). For succinyl-CoA, an aliquot was reconstituted in 50 mM ammonium acetate (pH 6.8 adjusted with ammonium hydroxide) and analysed as soon as possible once samples had been reconstituted to avoid degradation<sup>31-34</sup>. Liquid chromatography was performed with A = 50 mM ammonium acetate (pH 6.8) and B = 80% methanol. A Waters ACQUITY UPLC BEH C18 column (1.7 µm, 2.1 mm × 100 mm) was used at a flow rate of 250 µl/min. Initially, the mobile phase composition consisted of 2% B and held for 1.5 min after injection and its composition was increased to 15% over 1.5 min and then to 95% over 1.5 min and held for 9 min. The following quantifier and qualifier transitions (collision energy in eV) were used for succinyl-CoA: 868.1 → 99.0 (54 eV), 136.3 (54 eV), 259.1 (54 eV) and 361.3 (54 eV).

### Labile haem reporters targeted to subcellular compartments

HEK293T cells grown in DMEM with 10% FBS were transiently transfected in OptiMEM for 8 h using Dharmafect Duo transfection reagent (Dharmafect) in 96-well plate format. Peroxidase reporters (pEGFP-mitoAPX, pEGFP-APX, pEGFP-NLS-APX, and pmCherry-ER-HRP)<sup>17</sup> were co-transfected with 50 nM siRNA against *Pgrmc2*, *Pgrmc1*, the combination or a scramble control. After transfection, cells were switched to basal medium (DMEM with 10% FBS), basal medium plus 0.5 mM succinylacetone, haem-depleted medium (DMEM with 10% haem-depleted FBS) or haem-depleted medium plus 0.5 mM succinylacetone. Cells were lysed 72 h later in 100 µl haem lysis buffer (150 mM NaCl, 20 mM HEPES, 0.5% Triton X-100, with Protease Inhibitor Cocktail Set III). Fifty microlitres of lysate was incubated with the BioFX TMB One Component HRP microwell Substrate (Surmodics). Absorbance at 620 nm was measured after 5 min for the ER-HRP reporter, and after 30 min for mitochondrial, nuclear, and cytosolic APX reporters.

### Co-immunoprecipitation

Endogenous PGRMC2 and PGRMC1 were immunoprecipitated from primary brown adipocytes differentiated in vitro using anti-PGRMC2 and anti-PGRMC1 antibodies. Cells were lysed in IP lysis buffer (150 mM NaCl, 20 mM Tris-HCl, 10% glycerol, 1% Triton X-100 and complete EDTA-free protease inhibitor cocktail) and protein quantified using the DC assay (Biorad). One milligram of total proteome was incubated with 4 µg of anti-PGRMC2, anti-PGRMC1 or rabbit IgG control antibody pre-bound to 0.75 mg of Dynabeads Protein G (Thermo Fisher). After overnight incubation at 4 °C, beads-antibody-protein complexes were washed three times with PBS-0.02% Tween 20 for 5 min at RT, eluted in 50 mM glycine buffer pH = 2.8 for 10 min at 60 °C and separated by SDS-PAGE for immunodetection.

### Western blot analysis

Samples separated by SDS-PAGE were transferred onto nitrocellulose membranes. Membranes were incubated in blocking buffer (TBS-Tween 0.1%, BSA 5% w/v) for 1 h at room temperature. Membranes were incubated overnight at 4 °C with primary antibodies diluted in blocking buffer, washed three times for 15 min with TBS-Tween 0.1%, and incubated for 1 h at room temperature with HRP-conjugated secondary antibodies diluted in blocking buffer (1:20,000 dilution). The antibodies and dilutions used in this work were: PGRMC2 (1:1,000, Bethyl Laboratories, A302-954A and A302-955A), PGRMC1 (1:1,000, Bethyl Laboratories, A304-561A), PPAR $\gamma$ , EV-ERB $\alpha$  (1:200, Santa Cruz Biotechnology, sc-7273 and sc-100910), BACH1 (1:500, R&D Systems, AF5777), UCP1 and OxPhoS (1:5,000 and 1:300, Thermo Fisher Scientific, PA124894 and 458099), GAPDH, TUBULIN, and HSP90 (1:5,000, GeneTex, GTX627408, GTX27291, and GTX101423), and CEPB6 (1:1,000, Abgent, AP20492c).

### Primary adipocyte culture

Primary brown adipocytes were isolated from the interscapular BAT depot of wild-type and PATKO newborn mice. BAT depots were minced and digested by shaking for 40 min at 37 °C in isolation buffer containing 61.5 mM NaCl, 2.5 mM KCl, 0.65 mM CaCl<sub>2</sub>, 2.5 mM glucose, 50 mM HEPES, 50 U/ml, 50 µg/ml Pen/Strep, BSA 2% (w/v) and 1.5 mg/ml collagenase type I (Worthington). Cells were filtered through a 70-µm strainer and plated in DMEM with 25 mM glucose, 20 mM HEPES, 20% FBS and Pen/Strep. Differentiation was induced when cells reached confluence by switching the medium to DMEM, 10% FBS, 20 nM insulin, 1 nM triiodothyronine (T3), 0.5 mM 3-isobutyl-1-methylxanthine (IBMX) and 2 µg/ml dexamethasone (Dex). Two days later, medium was replaced with DMEM, 10% FBS, 20 nM insulin and 1 nM T3. On day 4 of differentiation, cells were treated with 0.5 mM succinylacetone, or switched to haem-depleted FBS, for bioenergetics and gene/protein expression studies and analysed at day 7. Exogenous haemin at a final concentration of 20 µM was added 48 h before bioenergetics studies were performed. On day 7 of differentiation, adipocytes were treated with vehicle or 100 nM noradrenaline for 2 h for gene expression studies. For complementation experiments, cells were infected with lentiviruses expressing mCherry, wild-type human PGRMC2, a human PGRMC2 haem-binding mutant (3×M; Y137F, K193A and Y194F) at day 0 of differentiation in the presence of 5 µg/ml polybrene. Rev-Erbα and BACH1 knockdown in mature adipocytes was performed as previously described<sup>35</sup>.

### Mitochondrial bioenergetics measurements

The oxygen consumption rate of adipocytes was measured on a Seahorse XFe96 instrument. Primary brown adipocytes differentiated in vitro were re-plated at day 5 of differentiation on gelatin-coated XFe96 plates at a density of 8,000 cells per well. Two days after plating, cells were equilibrated in serum-free DMEM (Sigma-Aldrich D5030) containing 25 mM glucose, 10 mM sodium pyruvate, 2 mM glutamine and 5 mM HEPES pH 7.4 for 1 h before a mitochondrial stress test was performed at day 7 consisting of 3 min cycles of mixing and 2 min cycles of measurements. Basal respiration rates were measured, followed by sequential injections of oligomycin (2 µM), FCCP (1 µM) and rotenone (2 µM) plus antimycin A (RAA, 2 µM). To measure the acute response to adrenergic signalling stimulators, compounds were injected using one of the ports after measurements of basal respiratory rates were complete. Freshly isolated BAT mitochondria (4 µg per well) were transferred onto XFe96 plates containing isolation buffer 2 (IB2 = 220 mM mannitol, 70 mM sucrose, 10 mM KH<sub>2</sub>PO<sub>4</sub>, 5 mM MgCl<sub>2</sub>, 1 mM EGTA, 0.5 mM ADP, 2 µM rotenone, 10 mM succinate, 0.2% BSA and 2 mM HEPES pH 7.4), and plates centrifuged at 2,000g for 20 min at 4 °C. Oxygen consumption rate was measured after sequential injections at final concentrations of 4 µM oligomycin, 4 µM FCCP and 4 µM antimycin A. Each cycle consisted of 30 s of mixing followed by 2.5 min of measurements.

### Quantitative PCR and RNA-seq

Total RNA was isolated from cells and tissues using the Direct-zol RNA MiniPrep Plus kit (Zymo Research). Taqman-based quantitative real-time PCR was performed using the SuperScript III Platinum One-Step qRT-PCR reagent (Thermo Fisher Scientific). Samples were run in triplicate as multiplexed reactions normalized to an internal control (36B4; acidic ribosomal phosphoprotein P0 mRNA). Sequences of primers and probes used are included in Supplementary Information.

For RNA-seq, total RNA was extracted from BAT of wild-type and PATKO mice at 30 °C using the Direct-zol RNA extraction kit (Zymo Research). PolyA<sup>+</sup> RNA was fragmented and prepared into strand-specific libraries using the Illumina True-seq stranded RNA kit (Illumina) and analysed on an Illumina HiSeq 2500 sequencer. Libraries were

sequenced using single-end 50-bp reads at a depth of 10–15 million reads per library. Single-end sequencing reads were mapped to the mouse reference genome (mm9, NCBI37) using STAR (version 2.3.0.c, default parameters). Only reads that aligned uniquely to a single genomic location were used for downstream analysis (MAPQ >10). Gene expression values were calculated for read counts on exons of annotated RefSeq genes using HOMER. DEGs were calculated with four replicates per condition using EdgeR, and a threshold of adjusted *P* value <0.05 was used to call DEGs. DEGs were used for pathway and Gene ontology functional enrichment analysis using Ingenuity Pathway Analysis (Qiagen) and Metascape<sup>36</sup> (<http://metascape.org>). Heat maps were generated using RStudio software (package 'gplots'). Pie charts and Circos plots were generated with Metascape and Adobe Illustrator. Data are available in GEO (GSE124621). Cell type-specific regulatory elements were download from the ENCODE SCREEN portal, using biosample 'C57BL/6 brown adipose tissue male adult 24 weeks'. BAT-specific enhancers as annotated by ENCODE (typically high DNase and H3K27ac signal but no H3K4me3 signal) were lifted over to mm9 using UCSC LiftOver and associated to genes by proximity (20 kb from TSS). Homer 4.9.1 was used to find enriched known and de novo motifs in enhancers associated to genes of interest.

### Mouse studies

All procedures were approved by the Institutional Animal Care and Use Committee of The Scripps Research Institute and conducted in accordance with relevant ethical regulations. To generate mice with adipose-specific deletion of *Pgrmc2*, mice with floxed *Pgrmc2* alleles<sup>37</sup> and backcrossed to the C57BL/6J background (NNT mutant) were crossed with an Adipoq-Cre strain<sup>38</sup> (JAX stock 010803). Similarly, mice with dual deletion of *Pgrmc1*<sup>39</sup> and *Pgrmc2* in adipose tissue were generated by crossing mice with floxed *Pgrmc1* and *Pgrmc2* alleles to the Adipoq-Cre strain. Floxed littermates without the *cre* transgene were used as controls and are referred to as wild type. Mice were born at room temperature and moved to 30 °C two weeks after weaning. Experiments were performed after a minimum of 4 weeks of acclimatization to 30 °C. Mice were kept on a 12-h light–dark cycle and fed standard chow breeder diet (5058, Picolab) or 60% HFD (D12492, Research Diets) as specified. Male and female mice were used in separate gender-matched experiments. No gender-specific differences were observed. For molecular characterization, mice were euthanized at or around ZT5, extensively perfused with ice-cold PBS, and tissues collected and immediately frozen in liquid nitrogen. For circadian time-course analysis, wild-type and PATKO mice (*n* = 3 per group, per time point) were euthanized every 4 h over a period of 24 h and tissues harvested as described above.

### Energy balance studies

Energy balance parameters were determined in a computer-controlled open-circuit system (Oxymax) that is part of an integrated Comprehensive Laboratory Animal Monitoring System (Columbus Instruments), as previously described<sup>40</sup>. Body temperature was monitored using a rectal probe (RET-3 probe, TH-5 Thermalert Monitoring Thermometer, Physitem) in cold exposure experiments, and by radiotelemetry in all other experiments. Radiotelemetry was enabled by surgically implanting a transmitter (TA10TA-F10; Data Sciences) into the peritoneal cavity, as previously described<sup>41</sup>. Male mice (20 weeks old) were allowed to recover for 14 d post-surgery and were then acclimated for 3 d to the experimental environment before measurements were taken. Data were recorded by placing a cage containing a mouse implanted with a transmitter on a receiver plate (RPC-1; DataScience). Data collection and offline analysis were performed using the DATAQUEST A.R.T. software (DataScience). To test the response to the β<sub>3</sub>-adrenergic receptor agonist CL316,243 (1 mg/kg) or an equivalent volume of PBS, was administered via intraperitoneal injection to 20-week-old male mice housed at thermoneutrality at ZT4.5. Oxygen consumption rate and activity levels were monitored using the CLAMS system.

## Exposure to cold

Experiments were performed on male and female mice 12–14 weeks of age. Mice were individually caged with minimal bedding and free access to food and water. To start the cold challenge, they were transferred to 4 °C, with controls remaining at 30 °C, and body temperature was monitored every 30 min for a total of 2.5 h. In that amount of time, all PATKO mice became severely hypothermic and all cold-exposed mice were euthanized. Cold challenge experiments started at or around ZT5 (11.00).

## Labile haem quantification

To purify nuclear and mitochondrial fractions from BAT and liver, one lobe of BAT or 100 mg of liver were dounce-homogenized in isolation buffer 1 (IB1 = 220 mM mannitol, 70 mM sucrose, 5 mM EGTA, and 50 mM MOPS pH 7.4). After centrifugation at 1,000g for 10 min at 4 °C, the nuclear pellet was passed through a 100-µm strainer and washed 5 times in IB1. Mitochondria in the supernatant were isolated from the cytosolic fraction after a second centrifugation at 9,500g for 10 min at 4 °C and washed twice with IB1. The nuclear and mitochondrial pellets were resuspended in 50 µl dH<sub>2</sub>O, sonicated and protein content quantified using the DC assay (Bio-Rad). Five microlitres of 25 nM apo-HRP was incubated in 384-well format with 10 µg in 5 µl of purified mitochondrial or nuclear protein lysates. After 5 min at room temperature, 40 µl of haem assay buffer (50 µM Amplex UltraRed, 0.02% H<sub>2</sub>O<sub>2</sub> in 0.1 M NaH<sub>2</sub>PO<sub>4</sub>/Na<sub>2</sub>HPO<sub>4</sub> buffer, pH 6) was added and fluorescence (ex.-em. 490/585) measured immediately for 15 min.

## Iron quantification

Frozen tissue (BAT, about 50 mg) was pulverized and lysed in 50 mM NaOH. Non-haem iron content was quantified using 200 µg of protein lysate and the ferrozine method as previously described<sup>42</sup>.

## Blood chemistry measurements

Blood samples were collected either from the retro-orbital plexus of anaesthetized mice, or by cardiac puncture after euthanasia. Plasma was separated using BD Microtrainer PST tubes with lithium heparin. Triglycerides and non-esterified free fatty acids were measured using the Serum Triglyceride Determination Kit (Sigma) and the HR Series NEFA-HR(2) kit (Wako). Norepinephrine levels were quantified using an ELISA kit (Abnova). Insulin levels were determined using an Ultra-Sensitive Rat ELISA Kit (Crystal Chem).

## Tissue lipid content

Frozen tissue (liver, about 30 mg) was pulverized and lysed in RIPA buffer. Triglycerides were quantified in 10 µl of tissue lysate using the EnzyChrome Triglyceride Assay kit (EGTA-200, Bioassay Systems). Triglyceride content was normalized to tissue weight.

## Treatment with CPAG-1

C57BL/6J male mice fed a 60 kcal% fat diet (D12492, Research Diets) were purchased from The Jackson Laboratory (DIO, JAX stock 380050) and kept in the same diet throughout the studies. DIO mice (>12 weeks of HFD, 20 weeks of age), randomized based on weight and fasting glycaemia, were dosed intraperitoneally with CPAG-1 every other day (45 mg/kg in a 2:3:1:4 DMSO:PEG40:ethanol:PBS vehicle solution). Weight and fasted glucose levels were monitored weekly. Mice were fasted for 16 h before analysis of basal blood chemistry parameters. At the conclusion of treatment, tissues were collected and snap-frozen for RNA extraction and western blot analysis or fixed for histological examination.

## Glucose and insulin tolerance tests

For glucose tolerance tests, mice were fasted for 6 h, and blood was collected from the tail vein before and at timed intervals after oral

gavage of glucose (1 g/kg). Plasma glucose was measured with a One-touch Ultra glucometer (Johnson & Johnson). For insulin tolerance tests, mice fasted for 4 h were injected intraperitoneally with insulin (0.4 U/kg; Novolin, Novo Nordisk). Glucose levels were determined before and at timed intervals after injection of insulin.

## Histology

Liver, and brown (BAT) and white (WAT) adipose tissue were fixed in Z-Fix (Anatech), dehydrated, embedded in paraffin, and 3-µm- (liver and BAT) or 10-µm- (WAT) thick sections stained with haematoxylin and eosin. Cell size was analysed using ImageJ software.

## Electron microscopy

BAT depots were collected and immediately placed in fixative buffer (2.5% paraformaldehyde, 3% glutaraldehyde, 0.02% picric acid in cacodylate buffer, pH 7.3) and stored at 4 °C for 72 h. Fixative buffer was refreshed after 48 h. Tissues were extensively washed in 0.1 M sodium cacodylate buffer (pH 7.3) prior to post-fix incubation in 2% OsO<sub>4</sub> in 0.1 M sodium cacodylate buffer for 4 h (buffer was refreshed after 2 h). Tissues were then washed in 0.1 M sodium cacodylate buffer (pH 7.3) followed by water. Tissues were dehydrated in a graded ethanol series and infiltrated and embedded in Spurr resin (Sigma-Aldrich). Thin sections were post-stained with 2% uranyl acetate followed by lead citrate and examined in a FEI Philips CM100 electron microscope at 80 KV. Images were taken using Radius 1.3 software with a Megaview G2 CCD Camera (EMSIS GmbH).

## Statistics

Results from in vitro assays and cell culture data are presented as mean ± s.d. Data generated in mouse studies are presented as mean ± s.e.m. The number of mice used in each experiment is indicated in the figure legends. Statistical analysis was performed on Prism software (GraphPad) using Student's *t*-test for comparisons between two groups, one-way ANOVA with multiple comparisons for assessment of more than two groups, and two-way ANOVA with multiple comparisons for repeated measurements. Comparisons among specific groups were done using post-tests as indicated in the figure legends.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

Source data tables are provided for Figs. 1–4 and Extended Data Figs. 1–8. Full scans of all western blots are shown in the Supplementary Information. RNA-seq data are available in the Gene Expression Omnibus under accession number GSE124621. All other data supporting the findings in this study are available from the corresponding author upon request.

30. Breitkopf, S. B. et al. A relative quantitative positive/negative ion switching method for untargeted lipidomics via high resolution LC-MS/MS from any biological source. *Metabolomics* **13**, 30 (2017).
31. Demoz, A., Garras, A., Asiedu, D. K., Nettelband, B. & Berge, R. K. Rapid method for the separation and detection of tissue short-chain coenzyme A esters by reversed-phase high-performance liquid chromatography. *J. Chromatogr. B* **667**, 148–152 (1995).
32. Li, Q., Zhang, S., Berthiaume, J. M., Simons, B. & Zhang, G. F. Novel approach in LC-MS/MS using MRM to generate a full profile of acyl-CoAs: discovery of acyl-dephospho-CoAs. *J. Lipid Res.* **55**, 592–602 (2014).
33. Liu, X. et al. High-resolution metabolomics with acyl-CoA profiling reveals widespread remodeling in response to diet. *Mol. Cell. Proteomics* **14**, 1489–1500 (2015).
34. Neubauer, S. et al. LC-MS/MS-based analysis of coenzyme A and short-chain acyl-coenzyme A thioesters. *Anal. Bioanal. Chem.* **407**, 6681–6688 (2015).
35. Isidor, M. S. et al. An siRNA-based method for efficient silencing of gene expression in mature brown adipocytes. *Adipocyte* **5**, 175–185 (2016).
36. Zhou, Y. et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **10**, 1523 (2019).

37. Clark, N. C. et al. Conditional ablation of progesterone receptor membrane component 2 causes female premature reproductive senescence. *Endocrinology* **158**, 640–651 (2017).
38. Eguchi, J. et al. Transcriptional control of adipose lipid handling by IRF4. *Cell Metab.* **13**, 249–259 (2011).
39. McCallum, M. L. et al. Conditional ablation of progesterone receptor membrane component 1 results in subfertility in the female and development of endometrial cysts. *Endocrinology* **157**, 3309–3319 (2016).
40. Kok, B. P. et al. Intestinal bitter taste receptor activation alters hormone secretion and imparts metabolic benefits. *Mol. Metab.* **16**, 76–87 (2018).
41. Sanchez-Alavez, M., Bortell, N., Galmozzi, A., Conti, B. & Marcondes, M. C. Reactive oxygen species scavenger *N*-acetyl cysteine reduces methamphetamine-induced hyperthermia without affecting motor activity in mice. *Temperature* **1**, 227–241 (2014).
42. Riemer, J., Hoepken, H. H., Czerwinska, H., Robinson, S. R. & Dringen, R. Colorimetric ferrozine-based assay for the quantitation of iron in cultured cells. *Anal. Biochem.* **331**, 370–375 (2004).

**Acknowledgements** We thank I. Hamza for labile haem reporter plasmids; A. Kralli, A. Saghatelian, P. Tontonoz, R. L. Wiseman, L. Gerace, J. Z. Long, N. Mitro and J. Hogenesch for critical input; M. R. Wood and T. Fassel for assistance with electron microscopy and N. Hah for help with RNA-seq studies. R.S. thanks the UCLA QCBio Collaboratory community directed by M. Pellegrini. This work was funded by NIH grants DK099810 and DK114785 (E.S. and B.F.C.),

DK121196 and S10OD016357 (E.S.), and OD016564 (J.K.P. and J.J.P.). B.P.K. and V.A. were supported by fellowships 15POST25100007 and 17POST33660833 from the American Heart Association.

**Author contributions** A.G. and E.S. conceived the project, designed research and analysed data. A.G. and B.P.K. performed in vivo experiments. A.G., C.G., V.A. and B.P.K. carried out cell-based assays. A.G. and J.Y.L. performed gene-expression and biochemical analyses. A.S.K. prepared PGRMC2 proteins. S.M. prepared apo-Rev-Erba protein. J.R.M.-B. and W.R.W. carried out mass spectrometry experiments. A.G. and R.S. performed bioinformatic analysis. C.G.P. synthesized CPAG-1. J.J.P. and J. K.P. provided *Pgrmc2* and *Pgrmc1* floxed mice. R.C.-C. and B.C. contributed to energy balance studies. L.A.S., D.K., C.G.P., G.S. and B.F.C. provided advice and reagents. A.G. and E.S. wrote the manuscript and integrated comments from the other authors.

**Competing interests** The authors declare no competing interests.

#### Additional information

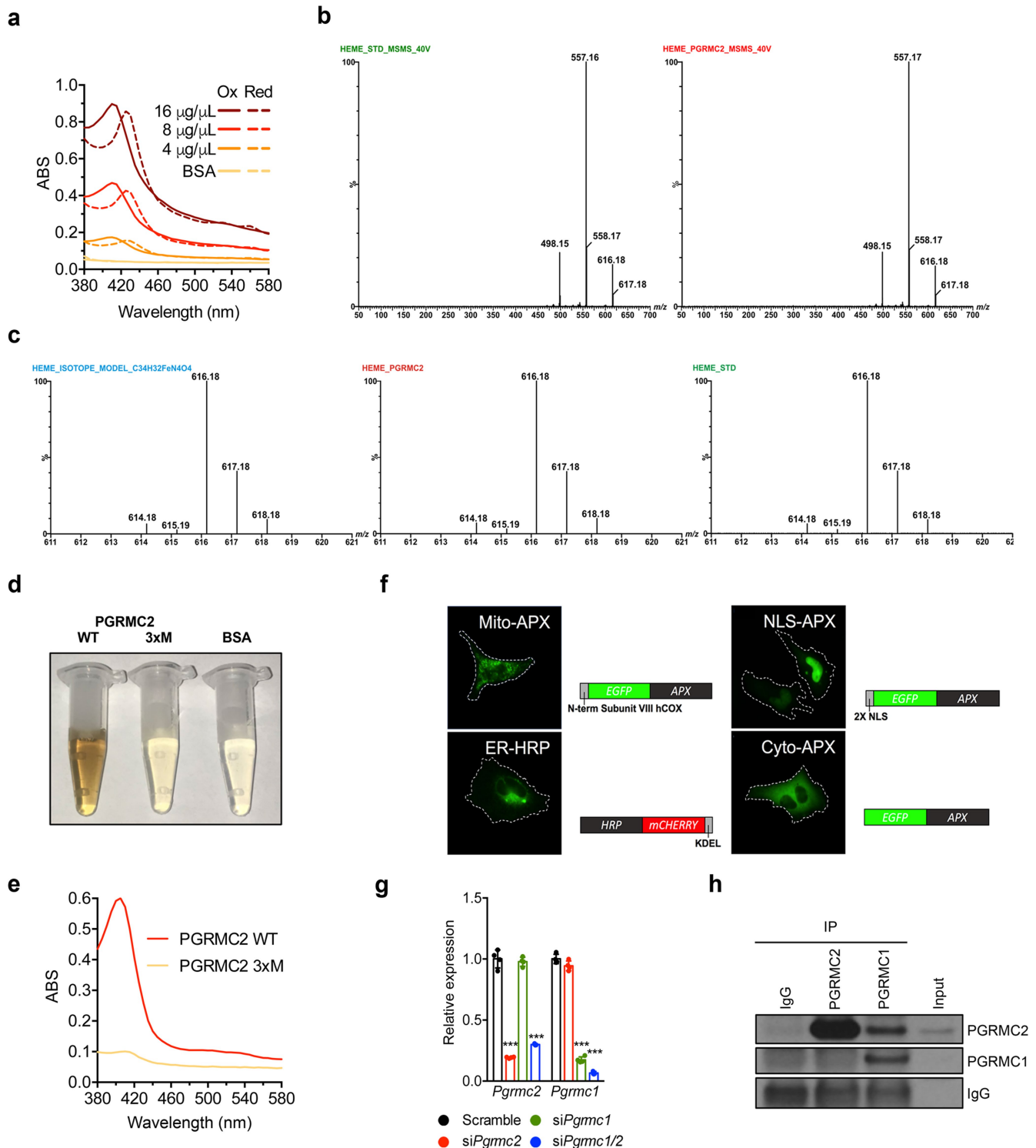
**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1774-2>.

**Correspondence and requests for materials** should be addressed to E.S.

**Peer review information** *Nature* thanks Urs Albrecht, Edward Chouchani, Iqbal Hamza and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

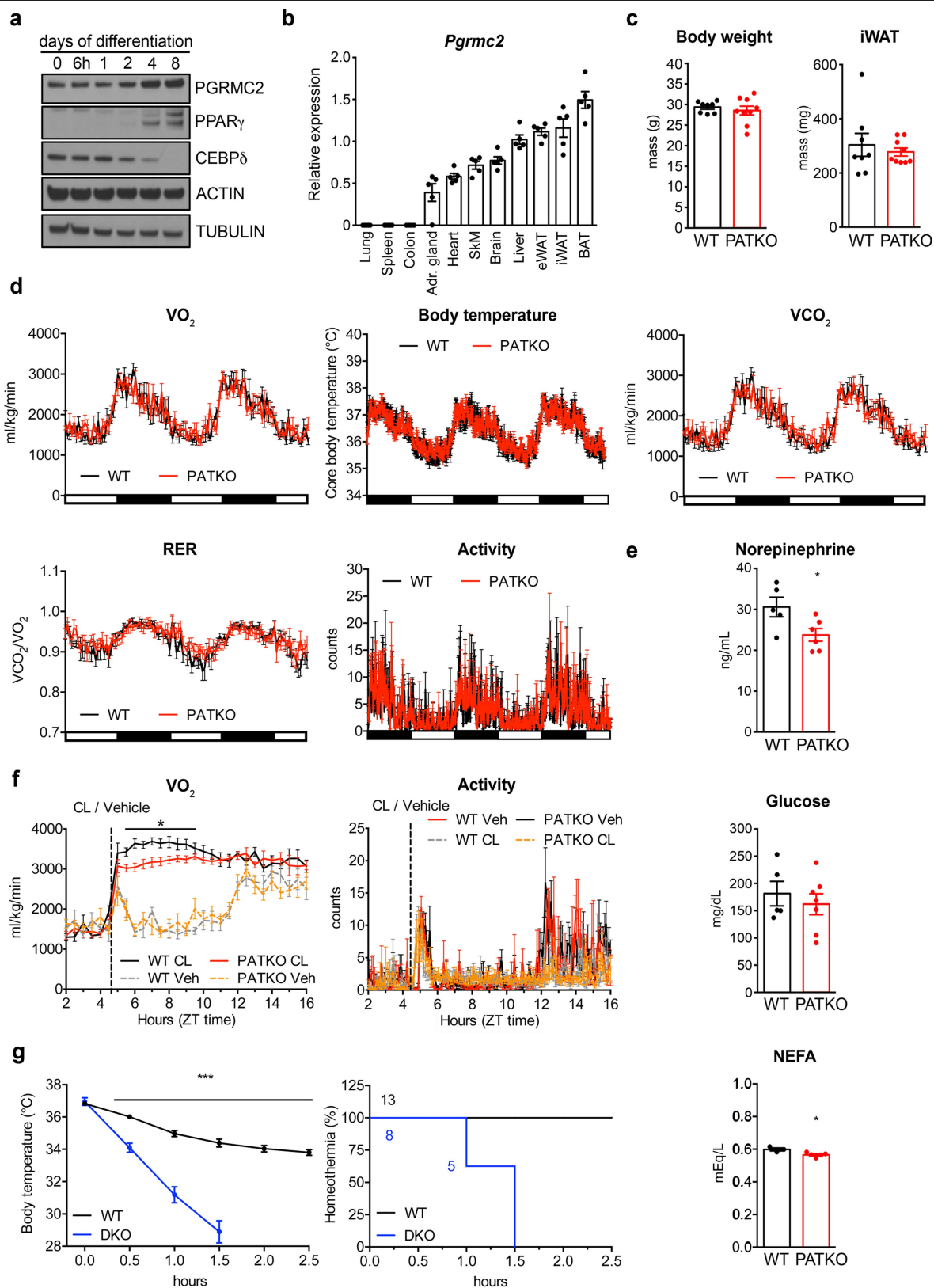
**Reprints and permissions information** is available at <http://www.nature.com/reprints>.





**Extended Data Fig. 1 | PGRMC2 binds haem and, with PGRMC1, coordinates its intracellular distribution.** **a**, Absorbance spectra of mouse PGRMC2 protein shows peaks of haem–protein complexes in the 390–450-nm range. Dotted spectra indicate haem–protein complexes after 10 mM dithionite reduction of the iron moiety. **b**, LC–MS/MS spectra of haemin standard (left) and PGRMC2 protein (right) with collision energy of 40 V. **c**, Isotope envelope of haemin calculated on the basis of isotope natural abundance for  $\text{C}_{34}\text{H}_{32}\text{ClFeN}_4\text{O}_4$  (left), PGRMC2 protein (centre) and haemin standard (right). **d**, Purified mouse PGRMC2(3xM) mutant (Y131F/K187A/Y188F) does not bind haem. **e**, The Soret peak typical of haemoproteins is absent in PGRMC2(3xM).

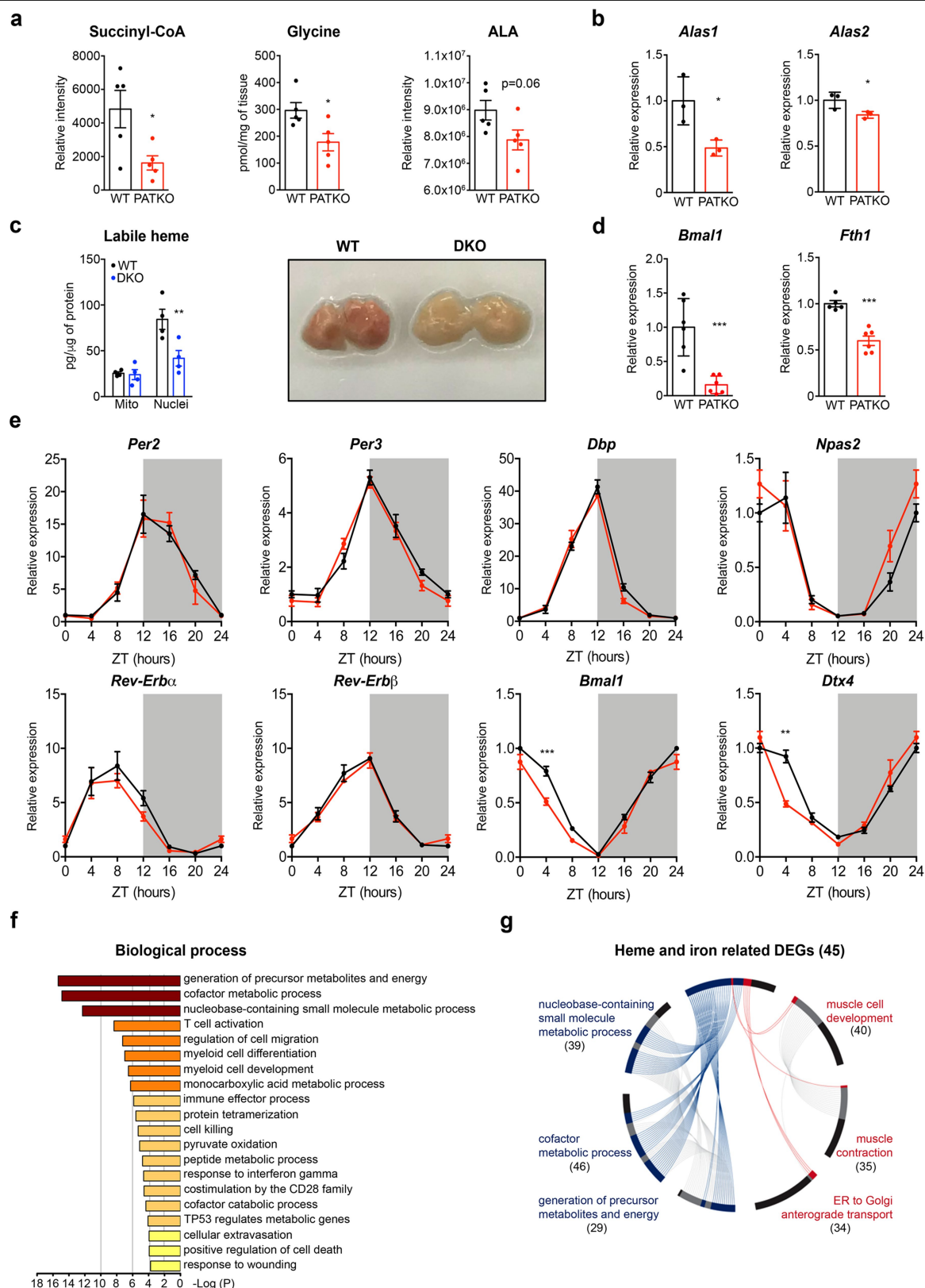
**f**, Representative fluorescence imaging of cells expressing targeted HRP or APX labile haem reporters, showing their localization to mitochondria, endoplasmic reticulum, nucleus and cytosol. **g**, Levels of *Pgrmc2* and *Pgrmc1* mRNA in siRNA-transfected HEK293T cells ( $n = 3$  biologically independent samples). **h**, Interaction of PGRMC1 with PGRMC2 is not observed when PGRMC2 is immunoprecipitated using an antibody that recognizes the haem-binding domain at the C terminus of PGRMC2. Representative results from two (**a–e**, **h**) or three (**f**, **g**) independent experiments. Data presented as mean  $\pm$  s.d., \*\*\* $P < 0.001$  versus scrambled basal; two-way ANOVA with multiple comparisons and a Tukey's post-test.



**Extended Data Fig. 2** | See next page for caption.

**Extended Data Fig. 2 | *Pgrmc2* is enriched in adipose tissue and regulates BAT function.** **a**, PGRMC2 protein levels increase during adipocyte differentiation. 3T3-L1 preadipocytes were induced to differentiate and protein extracts prepared at the indicated time points. PPAR $\gamma$  and CEBP $\delta$  are markers of mature adipocytes and preadipocytes, respectively. Representative results from three independent experiments. **b**, Profile of *Pgrmc2* mRNA expression across mouse tissues ( $n = 5$  biologically independent samples). **c**, Whole-body and inguinal subcutaneous fat weight of chow-fed wild-type and PATKO mice housed at 30 °C (WT,  $n = 8$ ; PATKO,  $n = 9$ ). **d**, OCR, core body temperature, CO<sub>2</sub> production rate, respiratory exchange ratio (RER), and activity oscillations of PATKO mice housed at 30 °C (WT,  $n = 5$ ; PATKO,  $n = 6$ ). **e**, Levels of plasma noradrenaline, glucose and non-esterified fatty acids

(NEFA) in wild-type and PATKO mice on cold challenge (WT,  $n = 5$ ; PATKO,  $n = 7$ ). **f**, Increased oxygen consumption upon acute injection of the  $\beta_3$ -agonist CL316,243 (1 mg kg<sup>-1</sup>) is reduced in PATKO mice housed at 30 °C, despite comparable motor activity ( $n = 5$  biologically independent samples). **g**, Adipose-specific PGRMC1 and PGRMC2 double-knockout mice (DKO) housed at 30 °C are cold-intolerant (WT,  $n = 13$ ; DKO,  $n = 8$  biologically independent samples). Survival curves of wild-type and DKO mice exposed to 4 °C (homeothermia is at 31 °C). Mice were exposed to 4 °C at ZT5. Data presented as mean  $\pm$  s.e.m. \* $P < 0.05$  and \*\*\* $P < 0.001$  versus wild type; two-tailed Student's  $t$ -test (**e**, **f**) or two-way ANOVA with multiple comparisons and a Tukey's post-test (**g**).



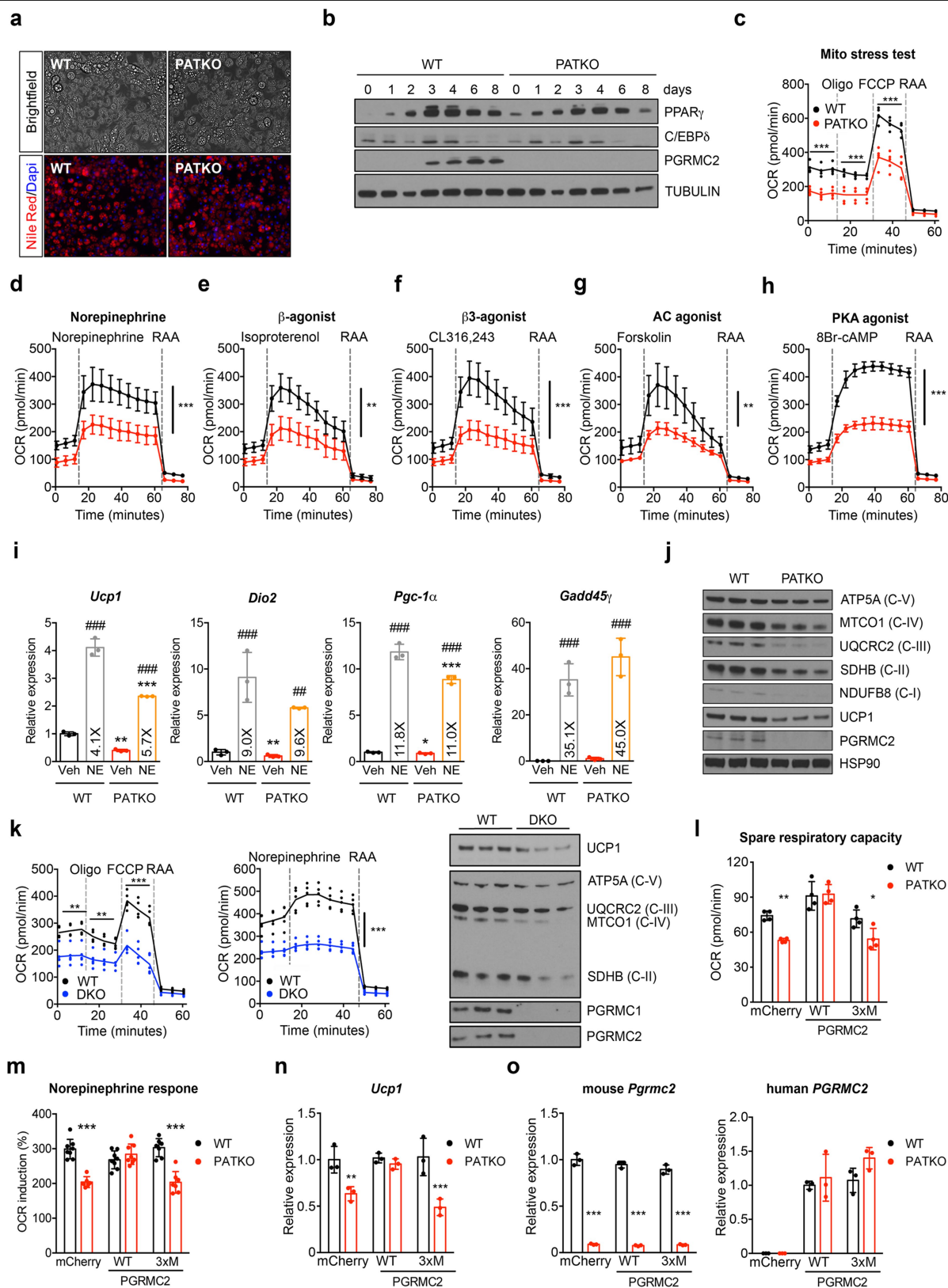
**Extended Data Fig. 3** | See next page for caption.

# Article

**Extended Data Fig. 3 | Effect of *Pgrmc2* deletion in BAT.** BAT from chow-fed wild-type and mutant mice housed at 30 °C was analysed. **a**, Levels of succinyl-CoA, glycine and aminolevulinic acid (ALA) in BAT quantified using targeted metabolomics ( $n=5$  biologically independent samples per group). **b**, PATKO mice show reduced expression of *Alas1* and *Alas2* in BAT ( $n=3$  biologically independent samples per group). **c**, Nuclear labile haem is significantly lower in BAT of fat-specific PGRMC1 and PGRMC2 DKO mice housed at 30 °C ( $n=4$  biologically independent samples per group). Similar to PATKO mice, BAT of DKO mice is discoloured. Representative results from two independent experiments. **d**, Expression of Rev-Erb $\alpha$  and BACH1 targets (*Bmal1* and *Fth1*, respectively) in BAT of PATKO mice housed at 30 °C (WT,  $n=5$ ; PATKO,  $n=6$ ). **e**, Circadian oscillation of clock components is not altered in PATKO BAT ( $n=3$

biologically independent samples per group per time point). **f**, Gene ontology (GO) category analysis (biological process) of significantly downregulated genes in RNA-seq analysis of BAT from wild-type and PATKO mice housed at 30 °C ( $n=4$  biologically independent samples per group). *P* values determined by standard accumulative hypergeometric statistical test. **g**, Circos plot of haem-related DEGs showing that the majority (28 out of 45) of them belong to the top-3 downregulated biological processes. Number in parentheses below each biological process represents the total number of DEGs in PATKO BAT in that category. Blue lines refer to downregulated DEGs and red lines to upregulated DEGs. Data presented as mean  $\pm$  s.e.m. \* $P<0.05$ , \*\* $P<0.01$  and \*\*\* $P<0.001$  versus wild type determined by two-tailed Student's *t*-test.



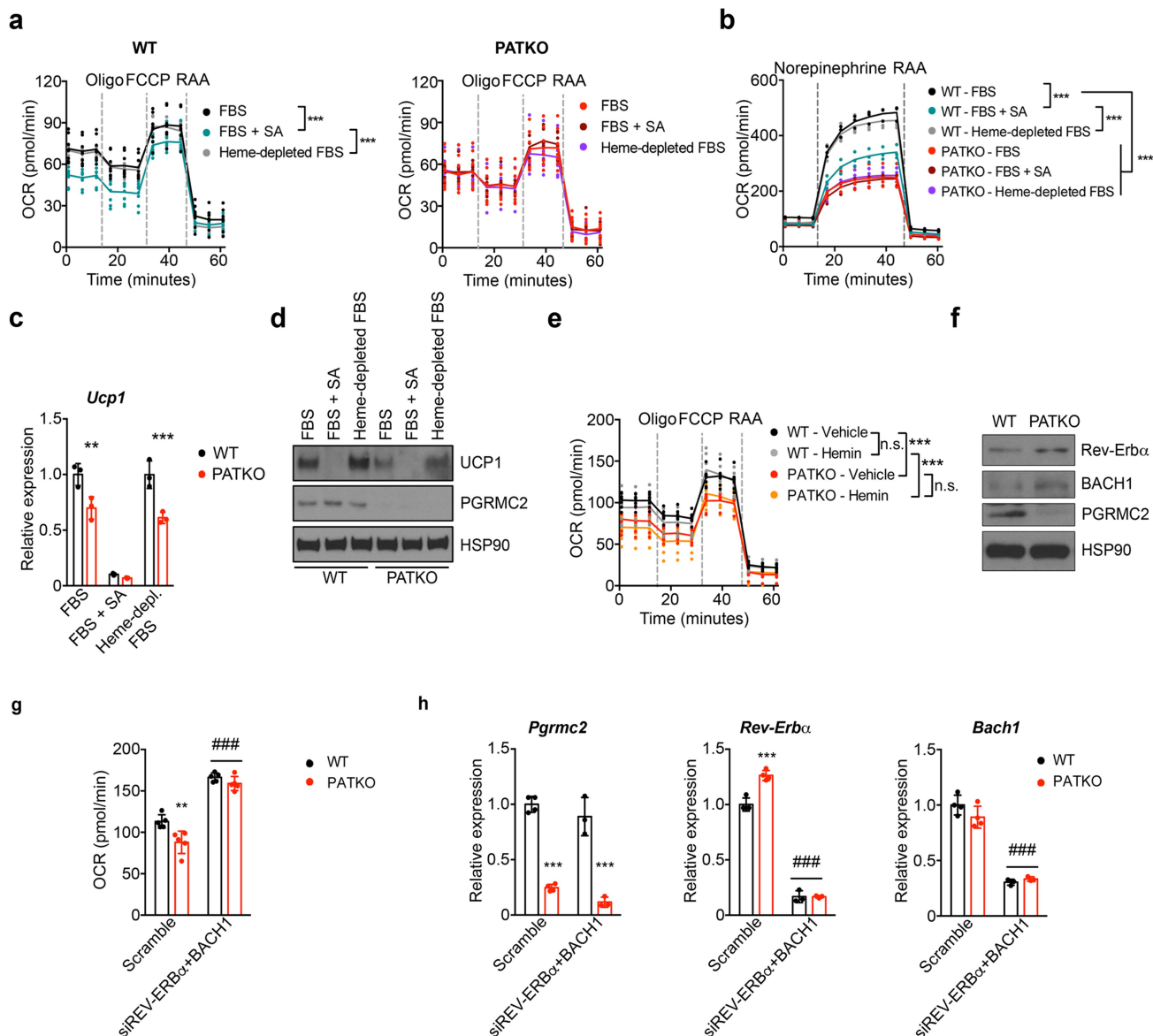


**Extended Data Fig. 4** | See next page for caption.

## Extended Data Fig. 4 | Primary brown adipocytes recapitulate the

**mitochondrial defects of PATKO BAT.** **a**, Wild-type and PGRMC2-null primary brown adipocytes differentiated in vitro imaged on day eight. Lipid stained with Nile red (red) and nuclei stained with Hoechst (blue). Scale bar, 100  $\mu$ m. **b**, Protein levels of adipocyte markers during the course of differentiation. **c**, PGRMC2-null brown adipocytes have impaired mitochondrial respiration ( $n = 3$ ). **d–h**, Lack of PGRMC2 in brown adipocytes results in a defective mitochondrial response to endogenous (**d**), synthetic pan  $\beta$ -adrenergic agonists (**e**) and pan  $\beta_3$ -adrenergic agonists (**f**), and to downstream activators of adrenergic signalling (**g**, **h**) ( $n = 5$ ). **i**, Induction of noradrenaline-responsive genes is similar in wild-type and PGRMC2-null brown adipocytes ( $n = 3$ ) exposed to 100 nM noradrenaline for 2 h. **j**, OXPHOS proteins and UCP1 are reduced in primary brown PATKO adipocytes. **k**, PGRMC1 and PGRMC2 DKO primary brown adipocytes differentiated in vitro show severe mitochondrial

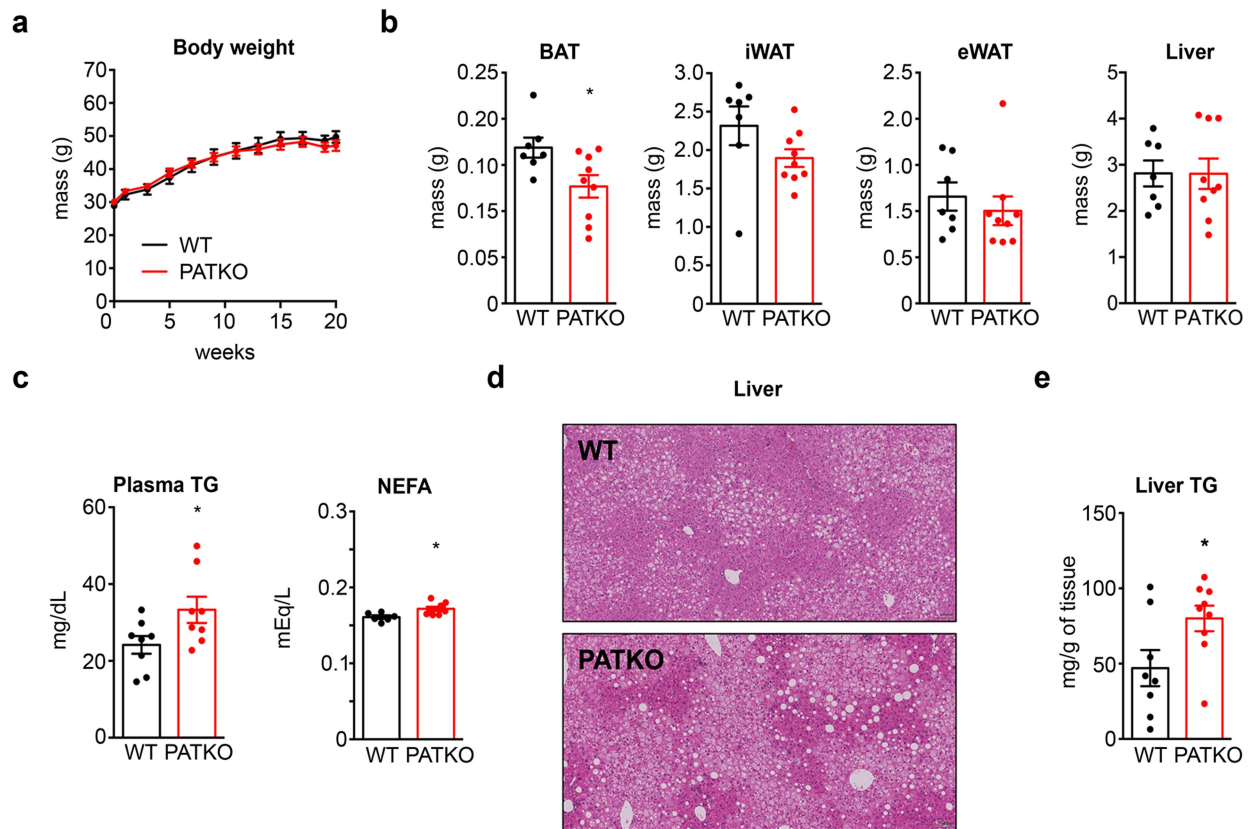
dysfunction, an inability to increase oxygen consumption on noradrenaline exposure ( $n = 3$ ), and reduced UCP1 and OXPHOS proteins. **l, m**, Overexpression of human wild-type PGRMC2, but not of a haem-binding mutant (3 $\times$ M (Y137F/K193A/Y194F)), can rescue mitochondrial function and the response to noradrenaline in PATKO adipocytes (**l**,  $n = 4$ ; **m**, WT-mCherry, WT-WT, PATKO-WT,  $n = 8$ ; WT-3 $\times$ M, PATKO-3 $\times$ M,  $n = 7$ ; PATKO-mCherry,  $n = 6$ ). **n**, *Ucp1* mRNA expression is restored when human wild-type PGRMC2, but not the haem-binding mutant 3 $\times$ M, is expressed in PATKO cells ( $n = 3$ ). **o**, Levels of mouse and human *Pgrmc2* mRNA in primary adipocytes used in **l–n** ( $n = 3$ ). In **a–o**,  $n$  represents biologically independent samples. Representative results from two (**j–o**) or three (**a–i**) independent experiments. Data presented as mean  $\pm$  s.d. \* $P < 0.05$ , \*\* $P < 0.01$  and \*\*\* $P < 0.001$  versus wild type; \*\*\* $P < 0.001$  versus vehicle; two-way ANOVA with multiple comparisons and a Bonferroni's post-test.



**Extended Data Fig. 5 | PGRMC2-mediated transport of endogenous labile haem regulates mitochondrial function in primary brown adipocytes.**

**a, b**, Inhibition for 48 h of endogenous haem synthesis with 0.5 mM succinylacetone (FBS + SA), but not exogenous haem depletion (haem-depleted FBS), in wild-type primary brown adipocytes phenocopies the mitochondrial defects of PATKO cells (**a**,  $n=8$ ; **b**,  $n=4$ ). **c, d**, Treatment with succinylacetone (0.5 mM) markedly reduces *Ucp1* mRNA and protein levels ( $n=3$ ). **e**, Exogenous haemin (20  $\mu$ M) does not correct mitochondrial dysfunction in PATKO cells ( $n=3$ ). **f**, PATKO brown adipocytes show higher

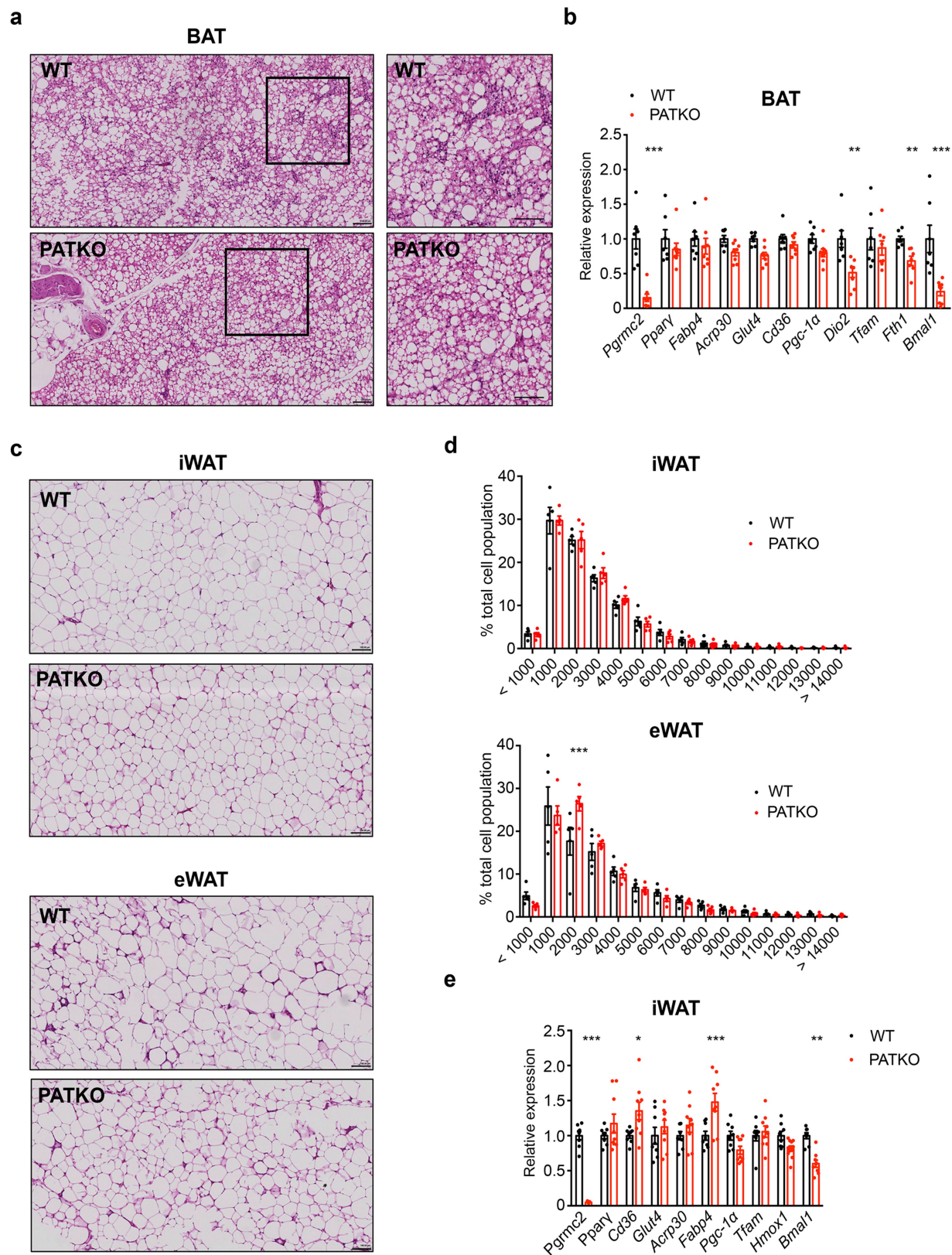
levels of Rev-Erb $\alpha$  and BACH1 protein. **g**, Dual knockdown of Rev-Erb $\alpha$  and BACH1 in mature PATKO adipocytes restores mitochondrial respiration ( $n=5$ ). **h**, *Pgrmc2*, *Rev-Erb $\alpha$*  (also known as *Nr1d1*) and *Bach1* mRNA in control and knockdown cells. In **a-h**,  $n$  represents biologically independent samples. Representative results from two independent experiments. Data presented as mean  $\pm$  s.d. \*\* $P < 0.01$  and \*\*\* $P < 0.001$  versus wild type; ### $P < 0.001$  versus scrambled; two-way ANOVA with multiple comparisons and a Bonferroni's post-test.



**Extended Data Fig. 6 | Body composition of PATKO mice fed a HFD.** Wild-type and PATKO mice were fed HFD for 20 weeks. **a**, Body weight progression (WT,  $n=7$ ; PATKO,  $n=9$ ). **b**, BAT of PATKO mice fed HFD is smaller compared to BAT of HFD-fed wild-type mice. No difference was seen in inguinal WAT (iWAT), epididymal WAT (eWAT) or liver weight (WT,  $n=7$ ; PATKO,  $n=9$ ). **c**, PATKO mice fed HFD had higher levels of plasma triglycerides and NEFA (WT,  $n=7$ ; PATKO,

$n=8$ ). **d**, H&E staining of liver shows increased steatosis in PATKO mice. Scale bar, 100  $\mu\text{m}$ . Representative images of seven biologically independent samples. **e**, PATKO mice fed HFD had more lipid accumulation in liver ( $n=8$ ). In **a–e**,  $n$  represents biologically independent samples. Data presented as mean  $\pm$  s.e.m.; \* $P < 0.05$  versus wild type; two-tailed Student's  $t$ -test.





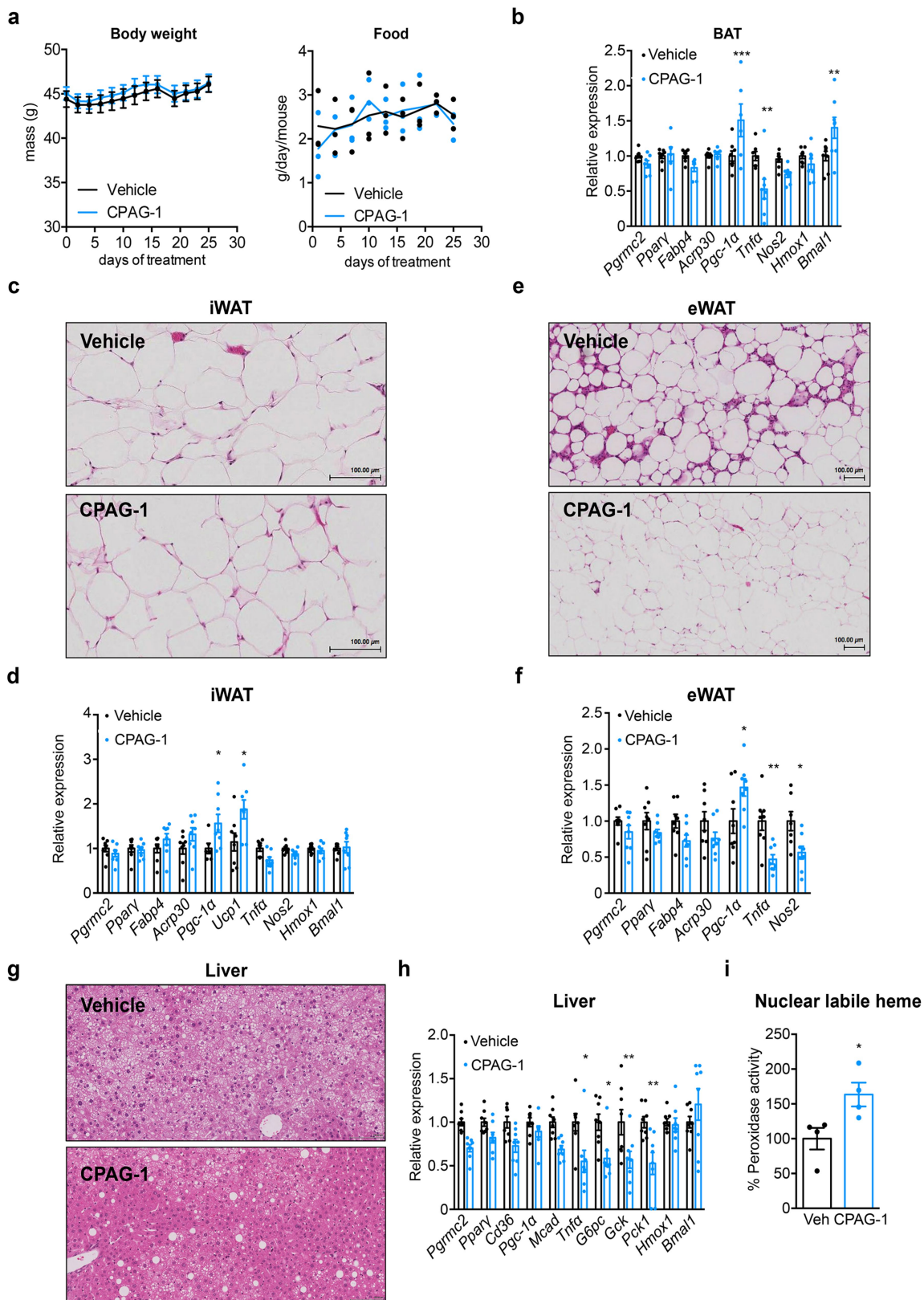
**Extended Data Fig. 7** | See next page for caption.



## Extended Data Fig. 7 | Analysis of adipose depots of PATKO mice fed a HFD.

Wild-type and PATKO mice were fed a HFD for 20 weeks. **a**, H&E stain images of BAT from wild-type and PATKO mice on a HFD show similar morphology. Insets are magnified on the right. Scale bar, 100  $\mu\text{m}$ . Representative images of seven biologically independent samples. **b**, Gene expression analysis in BAT shows reduced levels of *Fth1* and *Bmal1*, targets of BACH1 and Rev-Erb $\alpha$  respectively, in PATKO BAT (WT,  $n = 7$ ; PATKO,  $n = 8$ ). **c**, H&E staining of iWAT and eWAT from wild-type and PATKO mice fed HFD do not show clear differences. Scale bar, 100  $\mu\text{m}$ . Representative images of seven biologically independent samples.

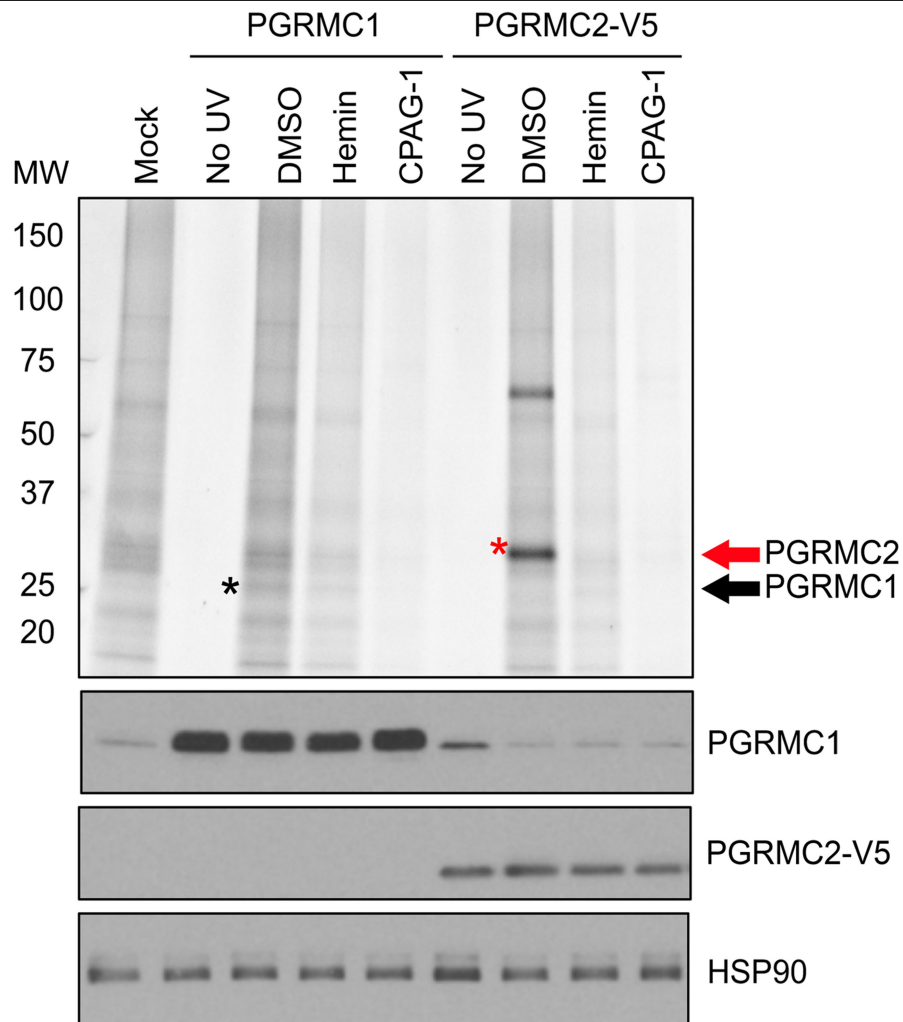
**d**, Size analysis of iWAT and eWAT adipocytes from HFD-fed wild-type and PATKO mice. The x axis indicates area in  $\mu\text{m}^2$  ( $n = 5$  images of biologically independent samples). **e**, Gene expression analysis in iWAT reveals a modest increase in expression of genes involved in lipid handling. Similar to BAT, *Bmal1* expression is significantly reduced in iWAT of PATKO mice (WT,  $n = 7$ ; PATKO,  $n = 9$ ). In **a–e**,  $n$  represents biologically independent samples. Data presented as mean  $\pm$  s.e.m.; \* $P < 0.05$ , \*\* $P < 0.01$  and \*\*\* $P < 0.001$  versus wild type; two-way ANOVA with multiple comparisons and a Bonferroni's post-test.



**Extended Data Fig. 8** | See next page for caption.

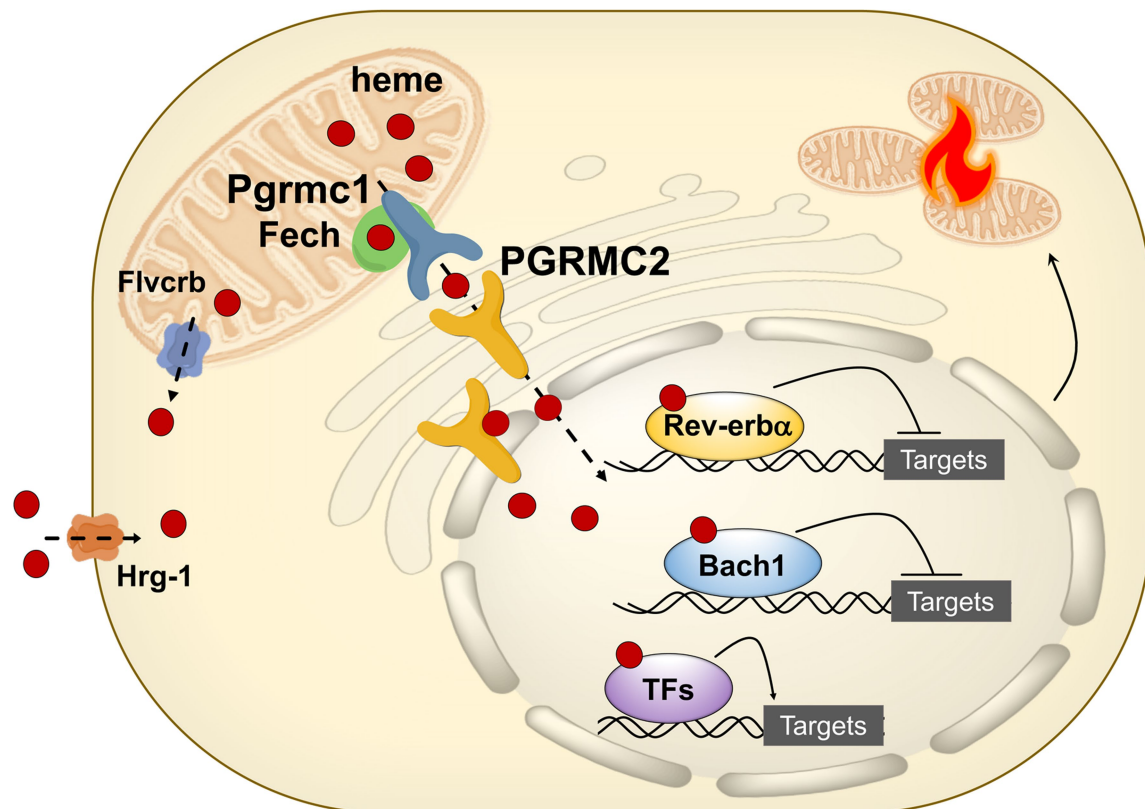
**Extended Data Fig. 8 | Effect of pharmacological activation of PGRMC2 in DIO mice.** DIO mice were treated with CPAG-1 for 30 days. **a**, Body weight (left) and food intake (right) progression ( $n = 8$ ). **b**, Expression of *Pgc-1 $\alpha$*  and *Bmal1* is increased in BAT of treated DIO mice ( $n = 8$ ). **c**, H&E staining of iWAT shows no difference between vehicle- and CPAG-1-treated DIO mice. Scale bar, 100  $\mu$ m. **d**, Gene expression analysis reveals increased expression of *Pgc-1 $\alpha$*  and *Ucp1* in iWAT of CPAG-1-treated DIO mice ( $n = 8$ ). **e**, H&E staining shows reduced fibrosis and immune cell infiltration in eWAT of DIO mice treated with CPAG-1. Scale bar, 100  $\mu$ m. **f**, Gene expression analysis shows decreased expression of markers of inflammation in eWAT of treated mice ( $n = 8$ ). **g**, H&E staining of liver shows that

CPAG-1 treatment modestly reduces lipid deposition. Scale bar, 100  $\mu$ m. **h**, Hepatic gene expression analysis shows decreased levels of gluconeogenic genes and inflammation markers in liver of treated mice ( $n = 8$ ). **i**, Treatment with CPAG-1 for four days significantly increases nuclear labile haem levels in the liver of DIO mice ( $n = 4$ ). In **a–i**,  $n$  represents biologically independent samples. Representative images of eight biologically independent samples per group (**d**, **e**, **g**). Data presented as mean  $\pm$  s.e.m.; \* $P < 0.05$ , \*\* $P < 0.01$  and \*\*\* $P < 0.001$  versus vehicle; two-way ANOVA with multiple comparisons and a Bonferroni's post-test.



**Extended Data Fig. 9 | Evaluation of interaction of CPAG-1 with PGRMC1 and PGRMC2 in live cells.** a, HEK293T cells transfected with expression vectors for either PGRMC1 or PGRMC2 were treated with 10  $\mu$ M probe 25 (the photoreactive form of CPAG-1) and DMSO, 100  $\mu$ M haemin or 100  $\mu$ M CPAG-1 for 30 min followed by UV photocross-linking, lysis and conjugation of labelled proteomes to a tetramethylrhodamine (TAMRA)-azide tag. Labelled proteomes were separated by SDS-PAGE and visualized by in-gel fluorescence

scanning. The intensity of the signals indicates the affinity of probe 25 for the overexpressed proteins. The black asterisk marks PGRMC1 protein and the red asterisk marks PGRMC2 protein. Although detectable, PGRMC1 shows very poor labelling with probe 25 relative to PGRMC2. Both interactions can be competed by haemin or CPAG-1. Western blot analysis confirms expression of PGRMC1 and PGRMC2 in transfected cells. Representative results from two independent experiments.



**Extended Data Fig. 10 | PGRMC2 is an intracellular haem chaperone critical for adipocyte function.** Model of the proposed role for PGRMC2 in haem dynamics in brown adipocytes. PGRMC2 acquires haem from PGRMC1, which forms a complex with FECH, the last enzyme in haem synthesis. PGRMC2, located in the endoplasmic reticulum and the nuclear envelope, facilitates delivery of labile haem to the nucleus. Nuclear labile haem alters expression of

genes regulated by haem-responsive transcriptional repressors such as Rev-Erbα and BACH1, which influence mitochondrial bioenergetics. FVLCR1b, a mitochondrial haem exporter identified in erythrocytes, and HRG-1, a plasma membrane haem importer characterized in macrophages, are also shown. FVLCR1b and HRG-1 are both expressed in brown adipocytes, but their role in haem dynamics in this cell type remains to be defined.



## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- ☐ ☒ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☒ ☐ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated
- ☐ ☒ Clearly defined error bars  
*State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on [statistics for biologists](#) may be useful.

### Software and code

Policy information about [availability of computer code](#)

Data collection

Datasets in this study were collected using the following software:

- SDS v2.4.1
- Wave v2.4.1.1
- SoftMax Pro v5.4.1
- STAR v2.3.0.c
- NIS-Elements AR v3.22.15
- ImageJ v1.48
- VitalView v5.1
- DATAQUEST A.R.T v4.0
- Radius v1.3
- MassLynx v4.1
- UCSC LiftOver ([https://genome.ucsc.edu/cgi-bin/hgLiftOver?hgslid=754627323\\_ju7bnkhsbYlo3wLsyVlmlbM1844A](https://genome.ucsc.edu/cgi-bin/hgLiftOver?hgslid=754627323_ju7bnkhsbYlo3wLsyVlmlbM1844A))

## Data analysis

RNAseq reads were mapped to the mouse genome mm9 NCBI37 using STAR 2.3.0.c (default parameters). Gene expression values were calculated using HOMER 4.9.1. DEGs were calculated with four replicates per group using EdgeR v3.5. DEGs were analyzed with Ingenuity Pathway Analysis (version 01-07, QIAGEN) Metascape (<http://metascape.org>), and HOMER (version 4.9.1, <http://homer.ucsd.edu/homer/motif/>). Additional software used in data analysis: GraphPad Prism v6.0h, RStudio v1.1.383, Image Lab v5.2.1, Microsoft Excel v16.28.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Source Data for Figs. 1-4 and Extended Data Figs. 1-8 are provided. Full scans for all western blots are provided in Supplementary Information. RNAseq data have been deposited in GEO under accession number GSE124621. All other data present in this study are available from the corresponding author upon request.

## Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](https://www.nature.com/authors/policies/ReportingSummary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were determined on the basis of previous experiments and account for biological/technical variability.
Data exclusions	No data were excluded from the analyses.
Replication	All data reported in this study were reproduced as biological replicates as stated in figure legends. All in vitro experiments were replicated at least twice. All in vivo experiments were replicated at least two independent cohorts.
Randomization	MS samples were processed in random order. For in vivo studies, DIO mice were randomly assigned to treatment groups based on weight and fasting glycemia.
Blinding	For MS analyses, experimenters were blinded to experimental conditions. Blinding was not possible in mouse studies due to the need to genotype or treat mice accordingly.

## Reporting for specific materials, systems and methods

### Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Unique biological materials
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Unique biological materials

Policy information about [availability of materials](#)

Obtaining unique materials All unique materials are available from the corresponding author upon reasonable request.

## Antibodies

### Antibodies used

The antibodies and dilutions used in this work were:  
 PGRMC2 (1:1,000, Bethyl Laboratories, A302-954A and A302-955A)  
 PGRMC1 (1:1,000, Bethyl Laboratories, A304-561A)  
 PPARg (1:200, Santa Cruz Biotechnology, sc-7273)  
 REV-ERBa (1:200, Santa Cruz Biotechnology, sc-100910)  
 BACH1 (1:500, R&D Systems, AF5777)  
 UCP1 (1:5,000, Thermo Fisher Scientific, PA124894)  
 OxPhoS (1:300, Thermo Fisher Scientific, 458099)  
 GAPDH (1:5,000, GeneTex, GTX627408)  
 TUBULIN (1:5,000, GeneTex, GTX27291)  
 HSP90 (1:5,000, GeneTex, GTX101423)  
 CEBPd (1:1,000, Abgent, AP20492c)  
 Rabbit IgG (Abcam, 37415)  
 Anti-rabbit IgG HRP-conjugated (1:10,000, Jackson immunoresearch, 211-035-109)  
 Anti-mouse IgG HRP-conjugated (1:20,000, Jackson immunoresearch, 315-035-045)  
 Anti-goat IgG HRP-conjugated (1:10,000, Jackson immunoresearch, 705-035-003)

### Validation

Antibodies were validated for the application and species used in this study by their manufacturers, and by the authors by using tissues derived from mouse null mutants. Validation data for the antibodies used can be found as follows:  
 PGRMC2 <https://www.bethyl.com/product/A302-954A/PGRMC2+Antibody>  
<https://www.bethyl.com/product/A302-955A/PGRMC2+Antibody>  
 PGRMC1 <https://www.bethyl.com/product/A304-561A/PGRMC1+Antibody>  
 PPARg <https://www.scbt.com/scbt/product/ppargamma-antibody-e-8>  
 REV-ERBa <https://www.scbt.com/scbt/product/rev-erbalph-antibody-rs-14?requestFrom=search>  
 BACH1 [https://www.rndsystems.com/products/mouse-bach1-antibody\\_af5777](https://www.rndsystems.com/products/mouse-bach1-antibody_af5777)  
 UCP1 <https://www.thermofisher.com/antibody/product/UCP1-Antibody-Polyclonal/PA1-24894>  
 OxPhoS <https://www.thermofisher.com/antibody/product/OxPhos-Rodent-WB-Antibody-clone-Cocktail-Cocktail/45-8099>  
 GAPDH <https://www.genetex.com/Product/Detail/GAPDH-antibody-GT239/GTX627408>  
 TUBULIN <https://www.genetex.com/Product?category=0&keyword=GTX27291&page=1>  
 HSP90 <https://www.genetex.com/Product?category=0&keyword=GTX101423&page=1>  
 CEBPd <https://www.abgent.com/products/AW5199-R-Cebpd-Antibody-Center>  
 Rabbit IgG <https://www.abcam.com/rabbit-igg-polyclonal-isotype-control-ab37415.html>  
 Anti-rabbit IgG HRP-conjugated <https://www.jacksonimmuno.com/catalog/products/211-035-109>  
 Anti-mouse IgG HRP-conjugated <https://www.jacksonimmuno.com/catalog/products/315-035-045>  
 Anti-goat IgG HRP-conjugated <https://www.jacksonimmuno.com/catalog/products/705-035-003>

## Eukaryotic cell lines

Policy information about [cell lines](#)

### Cell line source(s)

HEK293T cells were obtained from ATCC (CRL-3216).

### Authentication

Short-tandem repeat (STR) profiling.

### Mycoplasma contamination

Cells were routinely tested for mycoplasma (at least once every two months) and were always negative.

### Commonly misidentified lines (See [ICLAC](#) register)

No commonly misidentified cell lines were used.

## Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

### Laboratory animals

Mice with floxed *Pgrmc2* alleles and backcrossed to the C57BL/6J background (NNT mutant) were crossed with an Adipoq-CRE strain (JAX stock 010803) to generate mice with adipose-specific deletion of *Pgrmc2*. Floxed littermates without the CRE transgene were used as controls and are referred to as WT. Similarly, mice with dual deletion of *Pgrmc1* and *Pgrmc2* in adipose tissue were generated by crossing mice with floxed *Pgrmc1* and *Pgrmc2* alleles to the Adipoq-CRE strain. Mice were born and weaned at room temperature and moved to 30°C two weeks after weaning. Studies were performed in male and female mice. Primary brown preadipocytes were isolated from male and female pups (0-2 days old). C57BL/6 DIO mice were purchased from Taconic at 18 weeks of age.

### Wild animals

This study did not involve wild animals.



# Bile acid metabolites control T<sub>H</sub>17 and T<sub>reg</sub> cell differentiation

<https://doi.org/10.1038/s41586-019-1785-z>

Received: 24 October 2018

Accepted: 17 September 2019

Published online: 27 November 2019

Saiyu Hang<sup>1,12</sup>, Donggi Paik<sup>1,12</sup>, Lina Yao<sup>2</sup>, Eunha Kim<sup>1</sup>, Trinath Jamma<sup>3</sup>, Jingping Lu<sup>4</sup>, Soyoung Ha<sup>1</sup>, Brandon N. Nelson<sup>5</sup>, Samantha P. Kelly<sup>5</sup>, Lin Wu<sup>6</sup>, Ye Zheng<sup>7</sup>, Randy S. Longman<sup>8</sup>, Fraydoon Rastinejad<sup>4</sup>, A. Sloan Devlin<sup>2</sup>, Michael R. Krout<sup>5</sup>, Michael A. Fischbach<sup>9\*</sup>, Dan R. Littman<sup>6,10\*</sup> & Jun R. Huh<sup>1,11\*</sup>

Bile acids are abundant in the mammalian gut, where they undergo bacteria-mediated transformation to generate a large pool of bioactive molecules. Although bile acids are known to affect host metabolism, cancer progression and innate immunity, it is unknown whether they affect adaptive immune cells such as T helper cells that express IL-17a (T<sub>H</sub>17 cells) or regulatory T cells (T<sub>reg</sub> cells). Here we screen a library of bile acid metabolites and identify two distinct derivatives of lithocholic acid (LCA), 3-oxoLCA and isoalloLCA, as T cell regulators in mice. 3-OxoLCA inhibited the differentiation of T<sub>H</sub>17 cells by directly binding to the key transcription factor retinoid-related orphan receptor- $\gamma$ t (ROR $\gamma$ t) and isoalloLCA increased the differentiation of T<sub>reg</sub> cells through the production of mitochondrial reactive oxygen species (mitoROS), which led to increased expression of FOXP3. The isoalloLCA-mediated enhancement of T<sub>reg</sub> cell differentiation required an intronic *Foxp3* enhancer, the conserved noncoding sequence (CNS) 3; this represents a mode of action distinct from that of previously identified metabolites that increase T<sub>reg</sub> cell differentiation, which require CNS1. The administration of 3-oxoLCA and isoalloLCA to mice reduced T<sub>H</sub>17 cell differentiation and increased T<sub>reg</sub> cell differentiation, respectively, in the intestinal lamina propria. Our data suggest mechanisms through which bile acid metabolites control host immune responses, by directly modulating the balance of T<sub>H</sub>17 and T<sub>reg</sub> cells.

Bile acids are cholesterol-derived natural surfactants that are produced in the liver and secreted into the duodenum. They are critical for lipid digestion, antibacterial defence and glucose metabolism<sup>1</sup>. Although 95% of bile acids are re-absorbed through the terminal ileum of the small intestine and recirculated to the liver, bacteria transform hundreds of milligrams of bile acids to secondary bile acids with unique chemical structures<sup>2,3</sup>. In the healthy human gut, concentrations of secondary bile acids are in the hundreds of micromolar range<sup>2,4</sup>. Some bile acids disrupt cellular membranes owing to their hydrophobic nature<sup>5</sup>, whereas other bile acids protect the gut epithelium<sup>6</sup> and confer resistance to pathogens such as *Clostridium difficile*<sup>7</sup>. Bile acids also influence gut-associated inflammation, which suggests their potential to regulate gut mucosal immune cells<sup>8,9</sup>. The immune-modulatory effects of bile acids have mostly been studied in the context of innate immunity<sup>10–12</sup>. A recent study reported the cytotoxic effects of bile acids on gut-residing T cells<sup>13</sup>; however, whether these acids modulate T cell function directly has not been thoroughly examined. Since the identification of digoxin—a plant-derived molecule that contains a

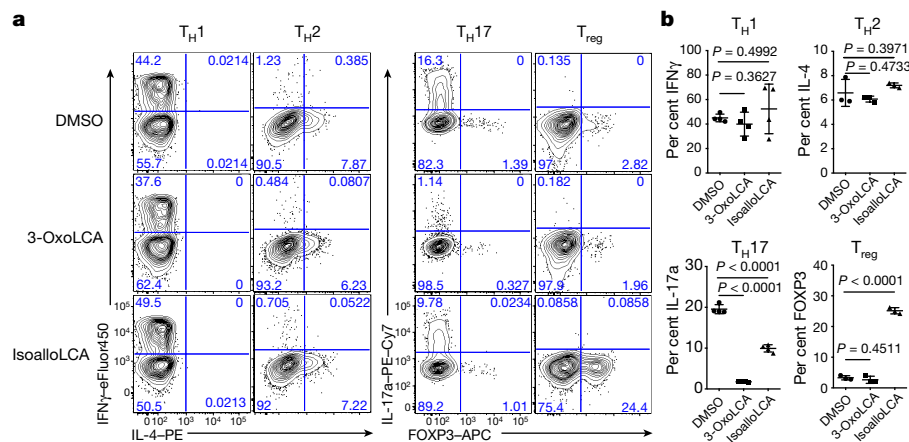
sterol-like core—as the first T<sub>H</sub>17 cell inhibitor that binds to ROR $\gamma$ t and inhibits its activity<sup>14</sup>, other structurally similar cholesterol derivatives have been identified as ROR $\gamma$ t modulators<sup>15–17</sup>. Because bile acids belong to a family of cholesterol metabolites and exist in the gut (where many T<sub>H</sub>17 cells reside<sup>18</sup>), we reasoned that bile acids control T<sub>H</sub>17 cell function by modulating ROR $\gamma$ t activity.

## Screen for T cell modulatory bile acids

To identify bile acids with modulatory effects on T cells, we screened about 30 compounds. Our screen included both primary bile acids that are synthesized by the host and secondary bile acids that are produced by bacterial modification of primary bile acids (Extended Data Fig. 1). Naive CD4<sup>+</sup> T cells were isolated from wild-type C57BL/6J (hereafter, B6Jax) mice and cultured with bile acids under T<sub>H</sub>17-cell differentiation conditions and, as a counter-screen, T<sub>reg</sub>-cell differentiation conditions (Extended Data Fig. 2). Notably, two derivatives of LCA were found to substantially affect the differentiation of T<sub>H</sub>17 and T<sub>reg</sub> cells. 3-OxoLCA

<sup>1</sup>Department of Immunology, Blavatnik Institute, Harvard Medical School, Boston, MA, USA. <sup>2</sup>Department of Biological Chemistry and Molecular Pharmacology, Blavatnik Institute, Harvard Medical School, Boston, MA, USA. <sup>3</sup>Department of Biological Sciences, Birla Institute of Technology and Science, Hyderabad, India. <sup>4</sup>Target Discovery Institute, Nuffield Department of Medicine, University of Oxford, Oxford, UK. <sup>5</sup>Department of Chemistry, Bucknell University, Lewisburg, PA, USA. <sup>6</sup>The Kimmel Center for Biology and Medicine of the Skirball Institute, New York University School of Medicine, New York, NY, USA. <sup>7</sup>Immunobiology and Microbial Pathogenesis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA, USA. <sup>8</sup>Jill Roberts Center for IBD, Weill Cornell Medicine, New York, NY, USA. <sup>9</sup>Department of Bioengineering, Stanford University, Stanford, CA, USA. <sup>10</sup>Howard Hughes Medical Institute, New York, NY, USA. <sup>11</sup>Evergrande Center for Immunologic Diseases, Harvard Medical School and Brigham and Women's Hospital, Boston, MA, USA. <sup>12</sup>These authors contributed equally: Saiyu Hang, Donggi Paik. \*e-mail: fischbach@fischbachgroup.org; Dan.Littman@med.nyu.edu; jun\_huh@hms.harvard.edu





**Fig. 1 | 3-OxoLCA inhibits T<sub>H</sub>17 cell differentiation and isoalloLCA enhances T<sub>reg</sub> cell differentiation. **a**, **b**, Flow cytometry and its quantification of intracellular staining for IFN $\gamma$  and IL-4, or IL-17a and FOXP3, in sorted naive T cells from wild-type B6Jax mice activated and expanded in the presence of mouse T<sub>H</sub>1, T<sub>H</sub>2, T<sub>H</sub>17 and T<sub>reg</sub> cell polarizing cytokines ( $n = 4, 3, 4$  and  $3$ ,**

respectively, biologically independent samples). A low concentration of TGF $\beta$  ( $0.01 \text{ ng ml}^{-1}$ ) was used for T<sub>reg</sub> cell culture. DMSO, 3-oxoLCA ( $20 \mu\text{M}$ ) or isoalloLCA ( $20 \mu\text{M}$ ) was added on day 0 and CD4<sup>+</sup> T cells were gated for analyses on day 3 for T<sub>H</sub>17 and T<sub>reg</sub> cells, and day 5 for T<sub>H</sub>1 and T<sub>H</sub>2 cells. Data are mean  $\pm$  s.d., by unpaired  $t$ -test with two-tailed  $P$  value.

inhibited T<sub>H</sub>17 cell differentiation, as shown by reduced expression of IL-17a, and isoalloLCA enhanced T<sub>reg</sub> cells, as shown by increased FOXP3 expression (Fig. 1a, b, Extended Data Fig. 2d, e). Although isoalloLCA strongly enhanced FOXP3 expression in the presence of low—but not high—concentrations of TGF $\beta$  (Fig. 1a, Extended Data Fig. 3a–c), the T<sub>reg</sub>-cell-enhancing activity of isoalloLCA required TGF $\beta$ , as shown by the fact that pretreatment of cells with anti-TGF $\beta$  antibody prevented FOXP3 enhancement (Extended Data Fig. 3d, e).

The modulatory effects of 3-oxoLCA on T<sub>H</sub>17 cells, and isoalloLCA on T<sub>reg</sub> cells, were specific to cell type; neither compound affected T cell differentiation into type 1 or type 2 T helper (T<sub>H</sub>1 or T<sub>H</sub>2) cells, as assessed by the expression of the cytokines IFN $\gamma$  and IL-4 and the transcription factors T-bet and GATA3 (Fig. 1a, b, Extended Data Fig. 3f, g). Although 3-oxoLCA did not affect T<sub>reg</sub> cells (Fig. 1b, Extended Data Fig. 2e), isoalloLCA reduced the differentiation of T<sub>H</sub>17 cells by about 50% without affecting ROR $\gamma$ t expression (Fig. 1a, b, Extended Data Fig. 3h). Both compounds exhibited dose-dependent effects (Extended Data Fig. 4a). 3-OxoLCA did not affect cell proliferation, whereas the addition of isoalloLCA to T cells led to reduced proliferation compared to a dimethylsulfoxide (DMSO) control (Extended Data Fig. 4b). Treatment with isoalloLCA did not impair cell viability (Extended Data Fig. 4c) or activation mediated by T cell receptor (TCR), as indicated by a similar expression of markers of TCR activation (such as CD25, CD69, NUR77 and CD44) between isoalloLCA and control treatments (Extended Data Fig. 4d). TCR activation promotes the enhancement of T<sub>reg</sub> cells by isoalloLCA; increasing TCR activation with higher concentrations of anti-CD3 resulted in stronger effects on FOXP3 expression, without affecting cell viability (Extended Data Fig. 4e, f).

### 3-OxoLCA inhibits T<sub>H</sub>17 differentiation

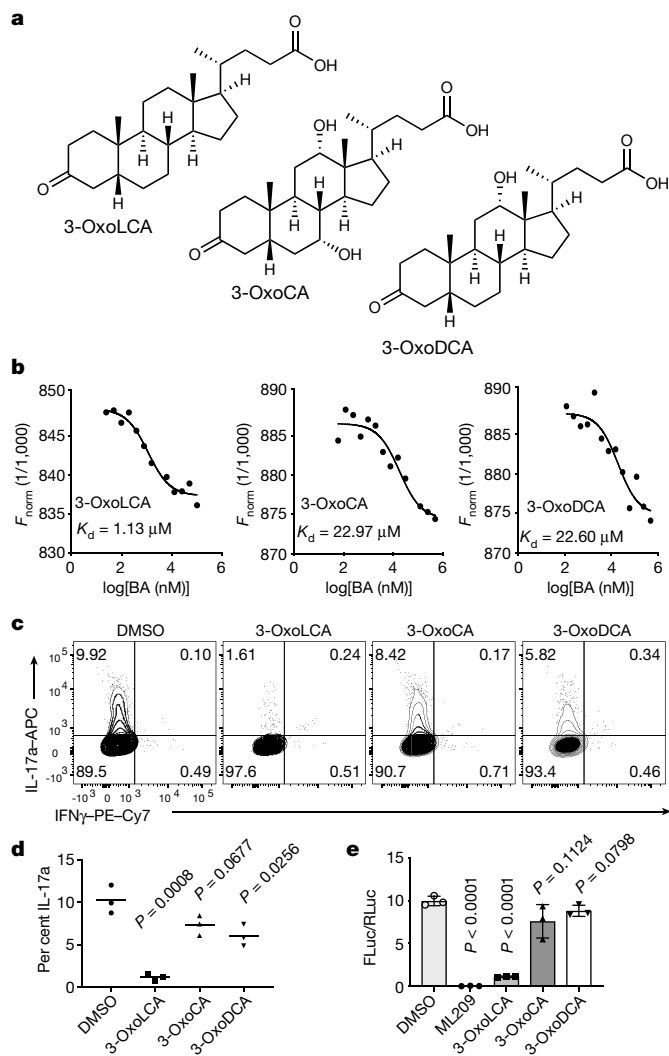
We next examined whether 3-oxoLCA physically interacts with the ROR $\gamma$ t protein in vitro. We performed a microscale thermophoresis assay using recombinant human ROR $\gamma$ t ligand-binding domain. 3-OxoLCA exhibited a robust physical interaction with the ROR $\gamma$ t ligand-binding domain at an equilibrium dissociation constant ( $K_d$ ) of about  $1 \mu\text{M}$ . We also tested two additional, structurally similar 3-oxo derivatives of bile acids, 3-oxocholeic acid (3-oxoCA) and 3-oxodeoxycholeic acid (3-oxoDCA) (Fig. 2a), and found that these derivatives had about 20 times higher  $K_d$  values than that of 3-oxoLCA (Fig. 2b). Neither 3-oxoCA nor 3-oxoDCA inhibited T<sub>H</sub>17 cell differentiation as robustly as did 3-oxoLCA (Fig. 2c, d). Next, we examined whether 3-oxoLCA

modulates the transcriptional activity of ROR $\gamma$ t. We assayed the effect of the bile acids on firefly luciferase expression directed by a fusion protein of ROR $\gamma$ t and GAL4 DNA-binding domain in human embryonic kidney (HEK) 293 cells<sup>14</sup>. Cells treated with ML209, a specific ROR $\gamma$ t antagonist, completely lost ROR $\gamma$ t activity<sup>19</sup>. Similarly, treatment with 3-oxoLCA significantly reduced the activity of the ROR $\gamma$ t reporter (Fig. 2e). These data suggest that 3-oxoLCA probably inhibits T<sub>H</sub>17 cell differentiation by physically interacting with ROR $\gamma$ t, and inhibiting its transcriptional activity.

### IsoalloLCA promotes T<sub>reg</sub> differentiation

We next sought to uncover the mechanism by which isoalloLCA exerts its enhancing effects on T<sub>reg</sub> cells. LCA has a  $3\alpha$ -hydroxyl group as well as a *cis* 5 $\beta$ -hydrogen configuration at the A–B ring junction and can undergo isomerization, presumably via the actions of gut bacterial enzymes<sup>2</sup>, to form isoLCA ( $3\beta,5\beta$ ), alloLCA ( $3\alpha,5\alpha$ ) or isoalloLCA ( $3\beta,5\alpha$ ) (Fig. 3a). Among these LCA isomers, isoalloLCA has the lowest log  $D$  value (2.2), comparable to the previously reported<sup>20,21</sup> log  $D$  values of chenodeoxycholic acid (2.2) and ursodeoxycholic acid (2.2) (Extended Data Table 1), which suggests that isoalloLCA is less lipophilic than the other isomers. IsoalloLCA, but not the other LCA isomers, enhanced FOXP3 expression, confirming that both the  $3\beta$ -hydroxyl group and *trans* (5 $\alpha$ -hydrogen) A–B ring configuration of isoalloLCA are required for enhancement of T<sub>reg</sub> cells (Fig. 3b). Compared to cells treated with DMSO, cells treated with isoalloLCA inhibited the proliferation of T effector cells in vitro, indicating they had acquired regulatory activity (Extended Data Fig. 5a, b). T cells isolated from FOXP3–GFP reporter mice exhibited both increased expression of *Foxp3* mRNA (Fig. 3c) and enhanced GFP levels after treatment with isoalloLCA (Extended Data Fig. 5c). Thus, the enhanced expression of FOXP3 induced by isoalloLCA occurs at the level of *Foxp3* mRNA transcription.

*Foxp3* transcription is regulated by three conserved non-coding enhancers known as CNS1, CNS2 and CNS3 (Fig. 3d), each of which has a distinct role in the development, stability and function of T<sub>reg</sub> cells<sup>22–24</sup>. Small molecules that promote T<sub>reg</sub> cells, such as the bacterial metabolite butyrate and the vitamin A derivative retinoic acid, enhance FOXP3 expression in a CNS1-dependent manner<sup>25,26</sup>. TGF $\beta$  also partially requires CNS1 for its T<sub>reg</sub>-cell-promoting activity, owing to the binding of its downstream signalling molecule SMAD3 to the CNS1 enhancer<sup>24,27</sup>. Whereas CD4<sup>+</sup> T cells from mice with deletions in CNS1 and CNS2 upregulated FOXP3 in response to isoalloLCA, cells



**Fig. 2 | 3-OxoLCA binds to ROR $\gamma$ t and inhibits its transcriptional activity.** **a**, Chemical structures of 3-oxoLCA, 3-oxoCA and 3-oxoDCA. **b**, Microscale thermophoresis assay. 3-OxoLCA binds to the ROR $\gamma$ t ligand-binding domain at a much lower  $K_d$  value than do the other two structurally similar bile acids. **c**, **d**, Flow cytometric analyses and quantification of IL-17a production from mouse naive CD4 $^{+}$  T cells cultured for 3 days under the T $_{H17}$  cell polarization condition ( $n = 3$  biologically independent samples per group). DMSO or bile acids at 20  $\mu$ M were added 18 h after cytokine addition. **e**, ROR $\gamma$ t luciferase reporter assay in HEK 293 cells treated with a positive control ML209 (2  $\mu$ M), 3-oxoLCA (10  $\mu$ M), 3-oxoCA (10  $\mu$ M), 3-oxoDCA (10  $\mu$ M) or DMSO. The ratio of firefly luciferase (FLuc) to *Renilla* luciferase (RLuc) activity is presented on the y axis ( $n = 3$  biologically independent samples per group). Data are mean  $\pm$  s.d., by unpaired *t*-test with two-tailed *P* value.

that lack CNS3 did not respond (Fig. 3e, f). By contrast, retinoic acid and TGF $\beta$  boosted T $_{reg}$  cell differentiation in CNS3-deficient cells, albeit with reduced efficiency (Extended Data Fig. 5d). Thus, unlike other small molecules that promote T $_{reg}$  cell differentiation (which do so in a CNS1-dependent manner), the FOXP3-enhancing activity of isoalloLCA requires CNS3.

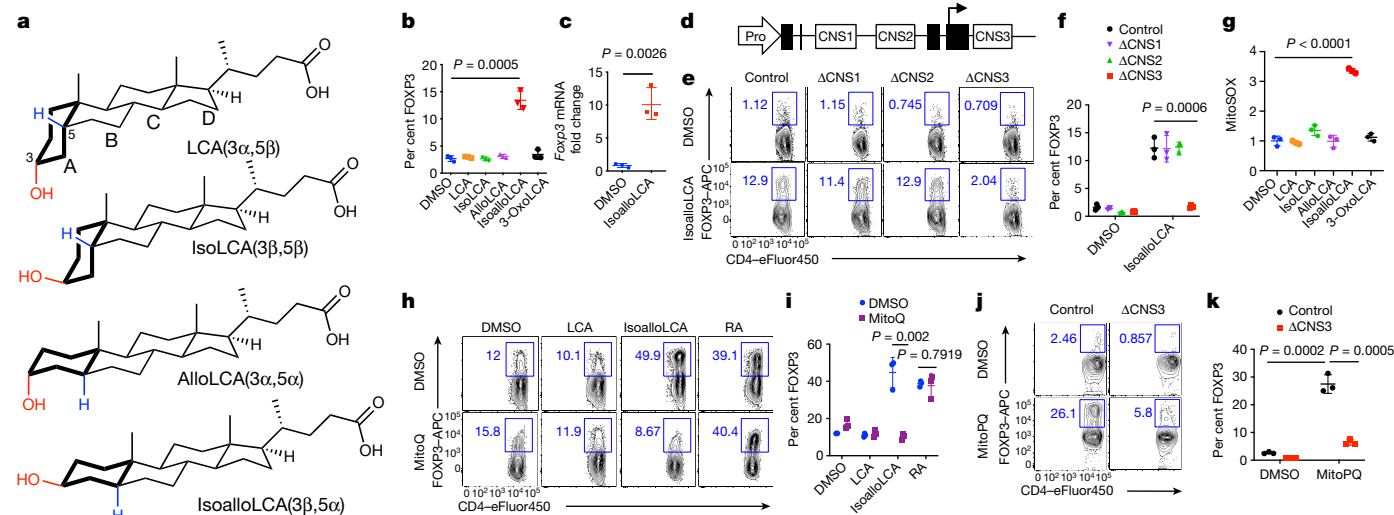
We investigated other known regulators of T cell function. The transcription factor REL binds to the CNS3 enhancer to induce FOXP3 expression<sup>24</sup>. We found that wild-type and REL-deficient cells express similar levels of FOXP3 upon treatment with isoalloLCA (Extended Data Fig. 5e, f). LCA targets the vitamin D receptor (VDR)<sup>28</sup> and the farnesoid X receptor (FXR)<sup>29</sup>. VDR has also previously been implicated in the modulation of both T $_{H17}$  and T $_{reg}$  cell function<sup>30–33</sup>. Compared

to a control treated with DMSO, isoalloLCA-treated cells deficient in VDR or FXR had similar amounts of FOXP3 induction (Extended Data Fig. 5g). Thus, the CNS3-dependent activation of FOXP3 by isoalloLCA is unlikely to be mediated through the actions of REL, VDR or FXR. VDR and FXR also did not contribute to the suppressive activities of 3-oxoLCA on T $_{H17}$  cells (Extended Data Fig. 5h). Of note, conjugating glycine to 3-oxoLCA or isoalloLCA reduced the immunomodulatory effects of these acids (Extended Data Fig. 5i–k).

CNS3 has previously been implicated in T $_{reg}$  cell development by promoting epigenetic modifications, such as H3K27 acetylation (H3K27ac) and H3K4 methylation, at the *Foxp3* promoter region<sup>23</sup>. Compared to cells treated with DMSO, cells treated with isoalloLCA had increased levels of H3K27ac in the *Foxp3* promoter region (Extended Data Fig. 6a). Consistent with this, treatment with isoalloLCA increased recruitment of the histone acetyltransferase p300 (Extended Data Fig. 6b). However, isoalloLCA did not affect H3K4 methylation (Extended Data Fig. 6c). The pan-bromodomain inhibitor iBET, which antagonizes H3K27ac, prevented the isoalloLCA-dependent enhancement of FOXP3 in a dose-dependent manner (Extended Data Fig. 6d, e). Consistent with previous work<sup>23</sup>, CNS3 deficiency not only reduced basal levels of H3K27ac but also abrogated the isoalloLCA-dependent increase in H3K27ac levels at the *Foxp3* promoter region (Extended Data Fig. 6f). Therefore, CNS3 is probably needed to establish a permissible chromatin landscape, whereupon the promoter region is further acetylated after treatment with isoalloLCA.

## MitoROS enhances FOXP3 expression

Cellular metabolism and epigenetic modification are intricately related; for example, the by-products of mitochondrial metabolism serve as substrates for histone acetylation and methylation<sup>34</sup>. T $_{reg}$  cells rely mainly on oxidative phosphorylation for their energy production<sup>35–37</sup>. Recent studies have identified two metabolites, 2-hydroxyglutarate and D-mannose, that promote T $_{reg}$  cell generation by modulating mitochondrial activities<sup>38,39</sup>. To assess whether isoalloLCA affects oxidative phosphorylation, we measured the oxygen consumption rate in T cells cultured for 48 h after treatment with DMSO or isoalloLCA. At this time point, FOXP3 is not yet strongly induced, which makes it possible to assess the effects of isoalloLCA on cellular metabolism before cells are fully committed to becoming T $_{reg}$  cells. Compared to DMSO, treatment with isoalloLCA increased the oxygen consumption rate in both wild-type and CNS3-knockout cells (Extended Data Fig. 6g), suggesting that treatment with isoalloLCA increases mitochondrial activity. Reactive oxygen species (ROS) are produced as by-products of mitochondrial oxidative phosphorylation. Whereas D-mannose increases cytoplasmic ROS production<sup>39</sup>, treatment with isoalloLCA led to increased production of mitoROS without affecting cytoplasmic ROS (Fig. 3g, Extended Data Fig. 6h, i). Unlike isoalloLCA, other isomers of LCA did not increase mitoROS production (Fig. 3g). Furthermore, isoalloLCA-treated cells displayed a modest, but significant increase in total mitochondrial mass and mitochondrial membrane potential (Extended Data Fig. 6j, k). To test whether mitoROS is directly involved in enhanced T $_{reg}$  cell differentiation by isoalloLCA, we used mitoQ (a mitochondrially targeted antioxidant) to reduce ROS levels in mitochondria (Extended Data Fig. 6l). Importantly, in the presence of mitoQ, isoalloLCA was no longer effective in enhancing T $_{reg}$  cell differentiation (Fig. 3h, i). By contrast, the retinoic-acid-dependent induction of T $_{reg}$  cells was unaffected by treatment with mitoQ (Fig. 3h, i). We next investigated whether mitoROS production is responsible for the enhanced levels of H3K27ac at the *Foxp3* promoter seen in cells treated with isoalloLCA. Co-treating cells with isoalloLCA and mitoQ decreased levels of H3K27ac, as compared to levels in cells treated with isoalloLCA only (Extended Data Fig. 6m). Stronger TCR stimulation that enhances FOXP3 expression (Extended Data Fig. 4d, e) also increased mitoROS production<sup>40</sup> (Extended Data Fig. 6n). Although TGF $\beta$ —which is essential for the FOXP3-enhancing



**Fig. 3 | MitoROS is necessary and sufficient for the isoalloLCA-dependent enhanced expression of FOXP3.** **a**, Chemical structures of LCA and its isomers: isoLCA, alloLCA and isoalloLCA. **b**, FOXP3 expression from mouse naive CD4<sup>+</sup> T cells cultured for 3 days with anti-CD3/28 and IL-2. DMSO or bile acids at 20  $\mu$ M were added to cell culture ( $n = 3$  biologically independent samples per group). **c**, Quantitative PCR (qPCR) analysis for *Foxp3* transcripts in cells treated with DMSO or isoalloLCA (20  $\mu$ M) ( $n = 3$  biologically independent samples per group). **d**, Diagram of the *Foxp3* gene locus containing the promoter region (Pro) and intronic enhancer regions (CNS1, CNS2 and CNS3). **e, f**, Flow cytometric analyses and quantification of CD4<sup>+</sup> T cells stained intracellularly for FOXP3. Naive CD4<sup>+</sup> T cells isolated from wild-type control, CNS1-, CNS2- or CNS3-knockout mice were cultured with anti-CD3/28 and IL-2, in the presence of DMSO or isoalloLCA (20  $\mu$ M) ( $n = 3$  biologically independent samples per group). **g**, MitoROS production measured by mitoSOX staining

with T cells cultured in the presence of DMSO or LCA isomers for 48 h. Staining intensity is reported as mean fluorescence intensity from flow cytometry analysis (PE channel). Different conditions were then normalized as the fold change relative to the values of DMSO condition ( $n = 3$  biologically independent samples per group). **h, i**, Representative fluorescence-activated cell sorting (FACS) plots and quantification of T cells stained intracellularly for FOXP3, cultured with anti-CD3/28, IL-2 and TGF $\beta$  (0.05 ng ml<sup>-1</sup>) in the presence of DMSO, LCA, isoalloLCA (20  $\mu$ M) or retinoic acid (RA) (1 nM), with DMSO or mitoQ (0.5  $\mu$ M) for 72 h ( $n = 3$  biologically independent samples per group). **j, k**, Flow cytometric analyses and quantification of CD4<sup>+</sup> T cells stained intracellularly for FOXP3. Naive CD4<sup>+</sup> T cells isolated from control or CNS3-knockout mice were cultured with anti-CD3/28 and IL-2 in the presence of DMSO or mitoPQ (10  $\mu$ M) ( $n = 3$  biologically independent samples per group). Data are mean  $\pm$  s.d., by unpaired *t*-test with two-tailed *P* value.

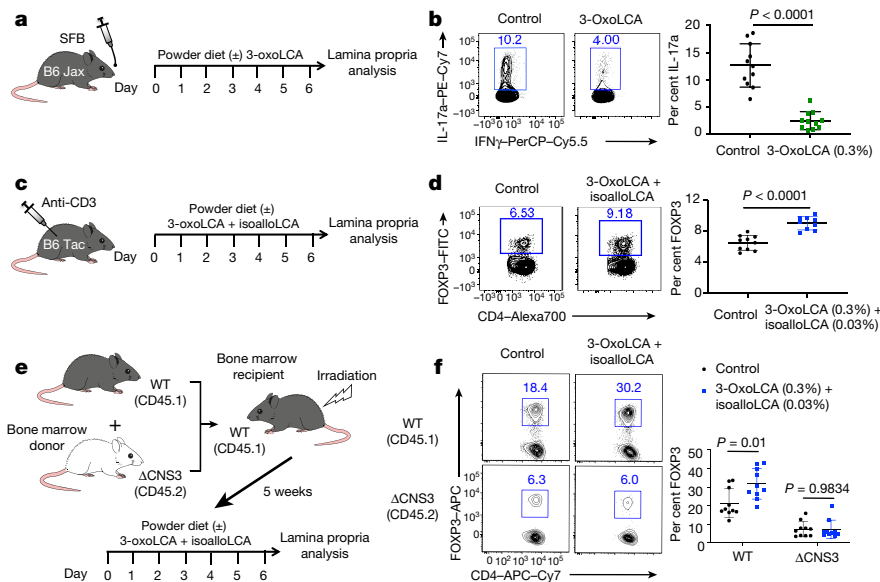
activity of isoalloLCA (Extended Data Fig. 3d, e)—was not required for increased mitoROS production (Extended Data Fig. 6o), it was required for the isoalloLCA-induced H3K27ac at the *Foxp3* promoter (Extended Data Fig. 6m). Because FOXP3 itself enhances mitochondrial oxidative phosphorylation<sup>41</sup>, and FOXP3-expressing T<sub>reg</sub> cells had higher levels of mitoROS than those of other subsets of CD4<sup>+</sup> T cells (Extended Data Fig. 6p), we investigated whether increased mitoROS production was a secondary effect of enhanced FOXP3 expression. CNS3-deficient cells that did not express high levels of FOXP3 in response to treatment with isoalloLCA nevertheless exhibited enhanced oxidative phosphorylation and increased levels of mitoROS (Fig. 3e, f, Extended Data Fig. 6g, q). We then investigated whether mitoROS is sufficient to promote T<sub>reg</sub> cell differentiation using the mitochondria-targeted redox cyclizer, mitoParaquat (mitoPQ)<sup>42</sup>. The addition of mitoPQ to a T cell culture was sufficient to enhance mitoROS production and T<sub>reg</sub> cell differentiation in a dose-dependent manner (Extended Data Fig. 6r, s). The T<sub>reg</sub> cell differentiation induced by mitoPQ, similar to that induced by isoalloLCA, required the CNS3 enhancer and TGF $\beta$  (Fig. 3j, k, Extended Data Fig. 6t). Together, our data support a model in which isoalloLCA promotes T<sub>reg</sub> cell differentiation by enhancing mitoROS production and increasing H3K27ac at the *Foxp3* promoter region, which also requires TGF $\beta$ -induced signalling (Extended Data Fig. 6u).

### Bile acids set T cell activities in vivo

We next examined whether 3-oxoLCA and isoalloLCA influence T<sub>H</sub>17 and T<sub>reg</sub> cell differentiation in vivo, using a mouse model. Segmented filamentous bacteria (SFB), a murine commensal, are known to induce T<sub>H</sub>17 cell differentiation in the small intestine of B6 mice<sup>43</sup>. C57BL/6NTac mice, from Taconic Biosciences (hereafter, B6 Tac), have abundant T<sub>H</sub>17 cells in their small intestine, owing to the presence of SFB. By contrast,

B6 Jax mice—which lack SFB—have few intestinal T<sub>H</sub>17 cells. To determine whether 3-oxoLCA suppresses T<sub>H</sub>17 cell differentiation in vivo, we gavaged B6 Jax mice with a faecal slurry containing SFB and fed these mice either a control diet or chow containing 0.3% (w/w) 3-oxoLCA for 1 week (Fig. 4a). The resulting average concentration of this metabolite in caecal contents was 24 pmol mg<sup>-1</sup> of wet mass (approximately equivalent to micromolar) (Extended Data Fig. 7a, b). This concentration was sufficient to suppress T<sub>H</sub>17 differentiation in vitro (Fig. 1c). Indeed, treatment with 3-oxoLCA significantly reduced the percentage of ileal T<sub>H</sub>17 cells (Fig. 4b). When we quantified the average levels of 3-oxoLCA in the stool of human patients with ulcerative colitis or in the caeca of conventionally housed mice, we observed a mean concentration of 23 or 1.0 pmol mg<sup>-1</sup>, respectively (Extended Data Fig. 7c, d). Levels of SFB colonization were comparable between control and 3-oxoLCA-treated groups of mice, which suggests that the change in the percentage of T<sub>H</sub>17 cells was not due to a decrease in SFB colonization (Extended Data Fig. 7e). In addition, B6 Tac mice (which have pre-existing SFB) had reduced percentages of T<sub>H</sub>17 cells when fed 3-oxoLCA, compared to when fed with vehicle (Extended Data Fig. 7f–h). Treatment with 3-oxoLCA did not affect percentages of T<sub>reg</sub> cells (Extended Data Fig. 7i). Even under the gut inflammatory conditions induced by anti-CD3 injection (which are known to produce a robust T<sub>H</sub>17 cell response<sup>18,44</sup>), mice treated with 1%—but not with 0.3%—3-oxoLCA had reduced the percentage of T<sub>H</sub>17 cells (Extended Data Fig. 7j–l).

To examine the effects of isoalloLCA on T<sub>reg</sub> cells in vivo, we fed SFB-colonized B6 Tac mice a control diet or a diet containing 0.03% (w/w) isoalloLCA. IsoalloLCA alone was insufficient to increase the percentage of T<sub>reg</sub> cells both at steady-state (Extended Data Fig. 7m) and after treatment with anti-CD3 (Extended Data Fig. 7n). We noted that 3-oxoLCA further enhanced the T<sub>reg</sub> cell differentiation induced by isoalloLCA in vitro (Extended Data Fig. 7o, p). Consistent with this



**Fig. 4 | 3-OxoLCA inhibits T<sub>H</sub>17 development and isoalloLCA enhances T<sub>reg</sub> cells in vivo.** **a, b,** Experimental scheme (**a**) and flow cytometric analysis (**b**) of T<sub>H</sub>17 cell induction by SFB. B6 Jax mice were gavaged with SFB-rich faecal pellets and kept on 3-oxoLCA (0.3%) for a week ( $n = 11$  mice per group). **c, d,** Experimental scheme (**c**) and flow cytometric analysis (**d**) of anti-CD3 experiment with a mixture of 3-oxoLCA + isoalloLCA ( $n = 10$  or 9 mice for control or 3-oxoLCA + isoalloLCA treatment, respectively). B6 Tac mice were intraperitoneally injected with anti-CD3 and fed a control diet or a mixture of 3-oxoLCA (0.3%) + isoalloLCA (0.03%) during the experiments.

observation, administration of a mixture of 0.3% (w/w) 3-oxoLCA and 0.03% (w/w) isoalloLCA significantly enhanced the T<sub>reg</sub> cell population in mice treated with anti-CD3, compared to that of mice with a control diet (Fig. 4c, d). Consistent with the mechanism in vitro, this treatment with the 3-oxoLCA and isoalloLCA mixture led to increased mitoROS production among CD4<sup>+</sup> T cells in the ileal lamina propria (Extended Data Fig. 7q). Importantly, the enhanced expression of FOXP3 induced by the 3-oxoLCA and isoalloLCA mixture in vivo was also dependent on the CNS3 enhancer, as shown by the fact that CNS3-knockout cells—unlike wild-type cells—no longer responded to this treatment in a mixed bone marrow experiment (Fig. 4e, f). Feeding both isoalloLCA and 3-oxoLCA in chow resulted in an average concentration of 47 pmol mg<sup>-1</sup> isoalloLCA in caecal contents (Extended Data Fig. 7b). This concentration was sufficient to enhance T<sub>reg</sub> cell differentiation in vitro (Fig. 1c). The mean concentration of isoalloLCA in the stool of patients with ulcerative colitis was 2 pmol mg<sup>-1</sup>, and ranged between 0 and 17 pmol mg<sup>-1</sup> (Extended Data Fig. 7c). These values are within an order of magnitude of the concentrations observed in mice fed 0.03% isoalloLCA and 0.3% 3-oxoLCA, suggesting that the in vivo levels of isoalloLCA achieved are physiologically relevant.

We next asked whether the immunomodulatory roles of 3-oxoLCA and isoalloLCA are mediated through changes in the composition of the gut bacterial community. 16S rDNA sequencing with faecal samples of mice fed diets containing bile acids revealed no substantial perturbations in the gut bacterial community, compared to mice fed a control diet (Extended Data Fig. 8a–e). Furthermore, treatment with 3-oxoLCA reduced T<sub>H</sub>17 cell induction in the colons of germ-free B6 mice infected with *Citrobacter rodentium* (Extended Data Fig. 8f, g). Thus, the T<sub>H</sub>17 and T<sub>reg</sub> cell modulatory activities of 3-oxoLCA and isoalloLCA probably do not require the presence of a community of commensal bacteria. These data suggest that both 3-oxoLCA and isoalloLCA directly modulate T<sub>H</sub>17 and T<sub>reg</sub> cell responses in mice in vivo.

Finally, we investigated whether the in vitro treatment of T cells with isoalloLCA produced T<sub>reg</sub> cells competent to exert suppressive function

**e, f,** Experimental scheme (**e**) and flow cytometric analysis (**f**) of T cells isolated from the ileal lamina propria. Bone marrow cells from wild-type (CD45.1) and CNS3-knockout (CD45.2) mice were mixed at a 1:1 ratio and transferred into irradiated wild-type (CD45.1) recipient mice. Five weeks after the transfer, recipient mice were fed a control diet or a diet containing a mixture of 3-oxoLCA (0.3%) + isoalloLCA (0.03%), followed by an anti-CD3 injection ( $n = 10$  mice per group). Data shown as mean  $\pm$  s.d. by unpaired  $t$ -test with two-tailed  $P$  value.

in vivo. The same number of FOXP3<sup>+</sup> T cells (CD45.2), sorted from T cell cultures with low or high TGF $\beta$  concentrations (TGF $\beta$ <sup>low</sup> or TGF $\beta$ <sup>high</sup> T<sub>reg</sub> cells, respectively) in the absence or presence of isoalloLCA, were adoptively transferred into RAG1-knockout mice that had also received CD45RB<sup>high</sup> naive CD4<sup>+</sup> T cells (CD45.1) (Extended Data Fig. 9a, b). Mice that received CD45RB<sup>high</sup> or CD45RB<sup>high</sup> and TGF $\beta$ <sup>low</sup> T<sub>reg</sub> cells developed substantial weight loss and shortened colon phenotypes, both of which are indicators of symptoms associated with colitis (Extended Data Fig. 9c–f). By contrast, the adoptive transfer of isoalloLCA-treated T<sub>reg</sub> cells protected mice from developing symptoms associated with colitis to the same degree as mice that received TGF $\beta$ <sup>high</sup> T<sub>reg</sub> cells (Extended Data Fig. 9c–f). T<sub>reg</sub> cells treated with isoalloLCA were more stable in terms of FOXP3 expression—as compared to TGF $\beta$ <sup>low</sup> T<sub>reg</sub> cells treated with DMSO—when analysed eight weeks after transfer (Extended Data Fig. 9g–j). In addition, mice that received isoalloLCA-treated T<sub>reg</sub> cells had reduced numbers of CD45.1<sup>+</sup> T effector cells (Extended Data Fig. 9k). Therefore, isoalloLCA probably promotes the stability of T<sub>reg</sub> cells and enhances their function after adoptive transfer in vivo, leading to decreased proliferation of T effector cells.

## Discussion

Some bile acids are thought to be tissue-damaging agents that promote inflammation, owing to their enhanced accumulation in patients with liver diseases and their chemical properties as detergents that disrupt cellular membranes<sup>45</sup>. However, recent studies have begun to reveal the anti-inflammatory roles of bile acids, particularly in the innate immune system by suppressing NF- $\kappa$ B-dependent signalling pathways<sup>46,47</sup> and by inhibiting NLRP3-dependent inflammasome activities<sup>41</sup>. Our studies reveal additional anti-inflammatory roles for two LCA metabolites found in both humans and rodents<sup>48–50</sup> that directly affect CD4<sup>+</sup> T cells: 3-oxoLCA suppresses T<sub>H</sub>17 cell differentiation and isoalloLCA enhances T<sub>reg</sub> cell differentiation. Our data suggest that both 3-oxoLCA and isoalloLCA are present in the stool samples of patients with colitis as well as in the



caeca of conventionally-housed B6 Jax mice (Extended Data Fig. 7c, d). Importantly, both bile acids are completely absent in germ-free B6 mice (Extended Data Fig. 7d). These data suggest that gut-residing bacteria may contribute to the production of 3-oxoLCA and isoalloLCA, although we cannot rule out the possibility that host enzymes are involved. Given the critical roles of  $T_H17$  and  $T_{reg}$  cells in a wide variety of inflammatory diseases and their close relationship with gut-residing bacteria, our study suggests the existence of novel modulatory pathways that regulate T cell function through bile acid metabolites. Future studies to elucidate the bacteria or host enzymes that generate 3-oxoLCA and isoalloLCA will provide the means for controlling T cell function in the context of autoimmune diseases and other inflammatory conditions.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1785-z>.

- Shapiro, H., Kolodziejczyk, A. A., Halstuch, D. & Elinav, E. Bile acids in glucose metabolism in health and disease. *J. Exp. Med.* **215**, 383–396 (2018).
- Ridlon, J. M., Kang, D. J. & Hylemon, P. B. Bile salt biotransformations by human intestinal bacteria. *J. Lipid Res.* **47**, 241–259 (2006).
- Devlin, A. S. & Fischbach, M. A. A biosynthetic pathway for a prominent class of microbiota-derived bile acids. *Nat. Chem. Biol.* **11**, 685–690 (2015).
- Hamilton, J. P. et al. Human cecal bile acids: concentration and spectrum. *Am. J. Physiol. Gastrointest. Liver Physiol.* **293**, G256–G263 (2007).
- Bernstein, H., Bernstein, C., Payne, C. M. & Dvorak, K. Bile acids as endogenous etiologic agents in gastrointestinal cancer. *World J. Gastroenterol.* **15**, 3329–3340 (2009).
- Barrasa, J. I., Olmo, N., Lizarbe, M. A. & Turnay, J. Bile acids in the colon, from healthy to cytotoxic molecules. *Toxicol. In Vitro* **27**, 964–977 (2013).
- Buffie, C. G. et al. Precision microbiome reconstitution restores bile acid mediated resistance to *Clostridium difficile*. *Nature* **517**, 205–208 (2015).
- Duboc, H. et al. Connecting dysbiosis, bile-acid dysmetabolism and gut inflammation in inflammatory bowel diseases. *Gut* **62**, 531–539 (2013).
- Martínez-Moya, P. et al. Dose-dependent antiinflammatory effect of ursodeoxycholic acid in experimental colitis. *Int. Immunopharmacol.* **15**, 372–380 (2013).
- Schaap, F. G., Trauner, M. & Jansen, P. L. Bile acid receptors as targets for drug development. *Nat. Rev. Gastroenterol. Hepatol.* **11**, 55–67 (2014).
- Guo, C. et al. Bile acids control inflammation and metabolic disorder through inhibition of NLRP3 inflammasome. *Immunity* **45**, 944 (2016).
- Ma, C. et al. Gut microbiome-mediated bile acid metabolism regulates liver cancer via NKT cells. *Science* **360**, eaan5931 (2018).
- Cao, W. et al. The xenobiotic transporter Mdr1 enforces T cell homeostasis in the presence of intestinal bile acids. *Immunity* **47**, 1182–1196 (2017).
- Huh, J. R. et al. Digoxin and its derivatives suppress  $T_H17$  cell differentiation by antagonizing ROR $\gamma$ t activity. *Nature* **472**, 486–490 (2011).
- Jin, L. et al. Structural basis for hydroxycholesterols as natural ligands of orphan nuclear receptor ROR $\gamma$ . *Mol. Endocrinol.* **24**, 923–929 (2010).
- Santori, F. R. et al. Identification of natural ROR $\gamma$  ligands that regulate the development of lymphoid cells. *Cell Metab.* **21**, 286–298 (2015).
- Soroosh, P. et al. Oxysterols are agonist ligands of ROR $\gamma$ t and drive Th17 cell differentiation. *Proc. Natl Acad. Sci. USA* **111**, 12163–12168 (2014).
- Esplugues, E. et al. Control of  $T_H17$  cells occurs in the small intestine. *Nature* **475**, 514–518 (2011).
- Huh, J. R. & Littman, D. R. Small molecule inhibitors of ROR $\gamma$ t: targeting Th17 cells and other applications. *Eur. J. Immunol.* **42**, 2232–2237 (2012).
- Roda, A., Minutello, A., Angellotti, M. A. & Fini, A. Bile acid structure-activity relationship: evaluation of bile acid lipophilicity using 1-octanol/water partition coefficient and reverse phase HPLC. *J. Lipid Res.* **31**, 1433–1443 (1990).
- Pellicciari, R. et al. Discovery of 3 $\alpha$ ,7 $\alpha$ ,11 $\beta$ -trihydroxy-6 $\alpha$ -ethyl-5 $\beta$ -cholan-24-oic acid (TC-100), a novel bile acid as potent and highly selective FXR agonist for enterohepatic disorders. *J. Med. Chem.* **59**, 9201–9214 (2016).

- Feng, Y. et al. Control of the inheritance of regulatory T cell identity by a cis element in the *Foxp3* locus. *Cell* **158**, 749–763 (2014).
- Feng, Y. et al. A mechanism for expansion of regulatory T-cell repertoire and its role in self-tolerance. *Nature* **528**, 132–136 (2015).
- Zheng, Y. et al. Role of conserved non-coding DNA elements in the *Foxp3* gene in regulatory T-cell fate. *Nature* **463**, 808–812 (2010).
- Arpaia, N. et al. Metabolites produced by commensal bacteria promote peripheral regulatory T-cell generation. *Nature* **504**, 451–455 (2013).
- Josefowicz, S. Z. et al. Extrathymically generated regulatory T cells control mucosal  $T_H17$  inflammation. *Nature* **482**, 395–399 (2012).
- Schlenner, S. M., Weigmann, B., Ruan, Q., Chen, Y. & von Boehmer, H. Smad3 binding to the foxp3 enhancer is dispensable for the development of regulatory T cells with the exception of the gut. *J. Exp. Med.* **209**, 1529–1535 (2012).
- Makishima, M. et al. Vitamin D receptor as an intestinal bile acid sensor. *Science* **296**, 1313–1316 (2002).
- Yu, J. et al. Lithocholic acid decreases expression of bile salt export pump through farnesoid X receptor antagonist activity. *J. Biol. Chem.* **277**, 31441–31447 (2002).
- Nandori, R. et al. The active form of vitamin D transcriptionally represses Smad7 Signaling and activates extracellular signal-regulated kinase (ERK) to inhibit the differentiation of a inflammatory T helper cell subset and suppress experimental autoimmune encephalomyelitis. *J. Biol. Chem.* **290**, 12222–12236 (2015).
- Jeffery, L. E. et al. 1,25-Dihydroxyvitamin D3 and IL-2 combine to inhibit T cell production of inflammatory cytokines and promote development of regulatory T cells expressing CTLA-4 and FoxP3. *J. Immunol.* **183**, 5458–5467 (2009).
- Gorman, S. et al. Topically applied 1,25-dihydroxyvitamin D3 enhances the suppressive activity of CD4<sup>+</sup>CD25<sup>+</sup> cells in the draining lymph nodes. *J. Immunol.* **179**, 6273–6283 (2007).
- Kang, S. W. et al. 1,25-Dihydroxyvitamin D3 promotes *FOXP3* expression via binding to vitamin D response elements in its conserved noncoding sequence region. *J. Immunol.* **188**, 5276–5282 (2012).
- Etcheberry, J. P. & Mostoslavsky, R. Interplay between metabolism and epigenetics: a nuclear adaptation to environmental changes. *Mol. Cell* **62**, 695–711 (2016).
- Gerriets, V. A. & Rathmell, J. C. Metabolic pathways in T cell fate and function. *Trends Immunol.* **33**, 168–173 (2012).
- Buck, M. D., O'Sullivan, D. & Pearce, E. L. T cell metabolism drives immunity. *J. Exp. Med.* **212**, 1345–1360 (2015).
- Gerriets, V. A. et al. Metabolic programming and PDHK1 control CD4<sup>+</sup> T cell subsets and inflammation. *J. Clin. Invest.* **125**, 194–207 (2015).
- Xu, T. et al. Metabolic control of  $T_H17$  and induced  $T_{reg}$  cell balance by an epigenetic mechanism. *Nature* **548**, 228–233 (2017).
- Zhang, D. et al. d-mannose induces regulatory T cells and suppresses immunopathology. *Nat. Med.* **23**, 1036–1045 (2017).
- Sena, L. A. et al. Mitochondria are required for antigen-specific T cell activation through reactive oxygen species signaling. *Immunity* **38**, 225–236 (2013).
- Angelin, A. et al. Foxp3 reprograms T cell metabolism to function in low-glucose, high-lactate environments. *Cell Metab.* **25**, 1282–1293 (2017).
- Robb, E. L. et al. Selective superoxide generation within mitochondria by the targeted redox cyclizer MitoParaquat. *Free Radic. Biol. Med.* **89**, 883–894 (2015).
- Ivanov, I. I. et al. Induction of intestinal Th17 cells by segmented filamentous bacteria. *Cell* **139**, 485–498 (2009).
- Gagliani, N. et al.  $T_H17$  cells transdifferentiate into regulatory T cells during resolution of inflammation. *Nature* **523**, 221–225 (2015).
- Trauner, M., Meier, P. J. & Boyer, J. L. Molecular pathogenesis of cholestasis. *N. Engl. J. Med.* **339**, 1217–1227 (1998).
- Vavassori, P., Mencarelli, A., Renga, B., Distrutti, E. & Fiorucci, S. The bile acid receptor FXR is a modulator of intestinal innate immunity. *J. Immunol.* **183**, 6251–6261 (2009).
- Pols, T. W. et al. TGR5 activation inhibits atherosclerosis by reducing macrophage inflammation and lipid loading. *Cell Metab.* **14**, 747–757 (2011).
- Kakiyama, G. et al. A simple and accurate HPLC method for fecal bile acid profile in healthy and cirrhotic subjects: validation by GC-MS and LC-MS. *J. Lipid Res.* **55**, 978–990 (2014).
- Sakai, K., Makino, T., Kawai, Y. & Mutai, M. Intestinal microflora and bile acids. Effect of bile acids on the distribution of microflora and bile acid in the digestive tract of the rat. *Microbiol. Immunol.* **24**, 187–196 (1980).
- Robben, J., Caenepeel, P., Van Eldere, J. & Eyssen, H. Effects of intestinal microbial bile salt sulfatase activity on bile salt kinetics in gnotobiotic rats. *Gastroenterology* **94**, 494–502 (1988).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019



## Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

### Mice

C57BL/6, FOXP3–GFP, FXR-knockout, VDR-knockout, CD45.1 and RAG1-knockout mice were purchased from the Jackson Laboratory. SFB-containing C57BL/6NTac mice were purchased from Taconic Bioscience. FOXP3–CNS-knockout and control mice were provided by the Y. Zheng laboratory. All mouse procedures were approved by the Institutional Animal Care and Use Committee at Harvard Medical School.

### Chemical synthesis of 3-oxoLCA, isoalloLCA, glyco-3-oxoLCA, and glyco-isoalloLCA

Detailed synthesis methods and characterization data are included in Supplementary Information.

### Measurement of lipophilicity

**Partitioning method.** Two microlitres of 10 mM stock solutions of each target compound was added to 1 ml each of 50 mM ammonium bicarbonate (pH = 8), and 1 ml of *n*-octanol in an Eppendorf tube. The resulting two-phase mixture was vortexed and then shaken for 18 h at 20 °C. The two phases were then carefully separated into the aqueous sample (bottom layer), and organic sample (top layer) and placed separately in autosampler vials and analysed by liquid chromatography with tandem mass spectrometry (LC–MS/MS).

**LC–MS method.** A Thermo q-Exactive Plus LC–MS equipped with an Ultimate 3000 HPLC was operated in negative ion mode after optimized to detect the  $[M - H]^-$  of the bile acids. Mobile phase A was 5 mM ammonium acetate with 0.012% formic acid and mobile phase B was HPLC-grade methanol. A Dikma Inspire C8 column (3- $\mu$ m particle size, 100-mm length, 4.6-mm inner diameter) was used for analysis. Each injection was 5  $\mu$ l, and a constant flow rate of 0.400 l/min was used. The gradient started at 0% B and was held constant for 2 min. Then, the mobile phase composition was linearly changed to 100% B over 8 min and held at 100% for the following 5 min. The mobile phase composition was changed to 0% B over the following 0.1 min, and the system was allowed to equilibrate to starting conditions over 1.9 min. Standards of all targets were used to establish retention times. Better-than-2-ppm mass accuracy was obtained on all measurements. The LC–MS analysis was done in triplicate, and the partitioning was done once for the  $[M - H]^-$  ion of each target.

**Calculation of log D.** The total response of target in octanol was divided by the total response in the aqueous phase to get a partitioning coefficient. Then,  $\log_{10}$  was taken and reported.

### Human faecal specimens

Faecal samples were obtained from patients with active ulcerative colitis under an Institutional-Review-Board-approved protocol and informed consent was obtained at Weill Cornell Medicine. Active inflammation was defined by a Mayo endoscopic score of > 1.

### In vivo bile acid analysis

Stock solutions of all bile acids were prepared by dissolving the compounds in molecular-biology-grade DMSO (Sigma Aldrich). These solutions were used to establish standard curves. Glycocholic acid or  $\beta$ -muricholic acid (Sigma Aldrich) was used as the internal standard for mouse and human samples, respectively. Bile acids were extracted from mouse caecal and human faecal samples, and quantified by ultra-high-performance liquid chromatography–mass spectrometry (UPLC–MS) as previously reported<sup>51</sup>. The limits of detection of individual bile

acids in tissues (in pmol/mg wet mass) are as follows:  $\beta$ -muricholic acid, 0.10; isoalloLCA, 0.45; isoLCA, 0.29; LCA, 0.12; alloLCA, 0.43; and 3-oxoLCA, 0.18.

### In vitro T cell culture

Naive CD4<sup>+</sup> (CD62L<sup>+</sup>CD44<sup>–</sup>CD25<sup>–</sup>CD4<sup>+</sup>) T cells were isolated from the spleens and the lymph nodes of mice of designated genotypes, using FACS. For some experiments, naive CD4<sup>+</sup> T cells were enriched using naive CD4<sup>+</sup> T cell isolation kits (Miltenyi). Naive CD4<sup>+</sup> T cells (40,000 cells) were cultured in a 96-well plate pre-coated with hamster IgG (MP Biomedicals) in T cell medium (RPMI, 10% fetal bovine serum, 25 mM glutamine, 55  $\mu$ M 2-mercaptoethanol, 100 U/ml penicillin, 100 mg/ml streptomycin) supplemented with 0.25  $\mu$ g/ml anti-CD3 (clone 145-2C11) and 1  $\mu$ g/ml anti-CD28 (clone 37.51). For naive T cell ( $T_H0$ ) culture, T cells were cultured with the addition of 100 U/ml of IL-2 (Peprotech). For  $T_H1$  cell differentiation, T cells were cultured with the addition of 100 U/ml of IL-2, 10  $\mu$ g/ml of anti-IL-4 (clone 11B11) and 10 ng/ml of IL-12 (Peprotech). For  $T_H2$  cell differentiation, T cells were cultured with the addition of 10  $\mu$ g/ml of anti-IFN $\gamma$  (clone XMGL2) and 10 ng/ml of IL-4 (R&D Systems). For  $T_H17$  cell differentiation, T cells were cultured with the addition of 10 ng/ml of IL-6 (eBioscience) and 0.5 ng/ml of TGF $\beta$  (Peprotech). For  $T_{reg}$  cell culture, T cells were cultured with the addition of 100 U/ml of IL-2 and various concentrations of TGF $\beta$ . For most in vitro experiments to test the effects of isoalloLCA, no additional TGF $\beta$  was added. Bile acids, retinoic acid (Sigma), mitoQ (Focus Biomolecules) or mitoPQ (Sigma) were added either at 0-h or 16-h time points. Compounds with low water solubility were sonicated before adding to the culture. Cells were collected and assayed by flow cytometry on day 3. For ROS and mitochondrial membrane potential detection, cells cultured for 2 days were incubated with 5  $\mu$ M of mitoSOX (ThermoFisher), 10  $\mu$ M of DCFDA (Sigma) or 2  $\mu$ M of JC-1 (ThermoFisher) for 30 min and assayed with flow cytometry.

### Flow cytometry

Cells collected from in vitro culture or in vivo mice experiments were stimulated with 50 ng/ml phorbol 12-myristate 13-acetate (PMA) (Sigma) and 1  $\mu$ M ionomycin (Sigma) in the presence of GolgiPlug (BD) for 4 h to determine cytokine expression. After stimulation, cells were stained with cell-surface marker antibodies and LIVE/DEAD Fixable dye, Aqua, to exclude dead cells, fixed and permeabilized with a FOXP3/transcription factor staining kit (eBioscience), followed by staining with cytokine- and/or transcription-factor-specific antibodies. All flow cytometry analyses were performed on an LSR II flow cytometer (BD) and data were analysed with FlowJo software (TreeStar).

### Cell proliferation assay

Naive CD4<sup>+</sup> T cells were labelled with 1  $\mu$ M carboxyfluorescein succinimidyl ester (CFSE, BioLegend) and cultured for three days before FACS analysis.

### In vitro suppression assay

A total of  $2.5 \times 10^4$  freshly purified naive CD4<sup>+</sup>CD25<sup>–</sup>CD44<sup>–</sup>CD62L<sup>high</sup> T ( $T_{conv}$ ) cells from CD45.1 B6 mice were labelled with 1  $\mu$ M CFSE, activated with soluble anti-CD3 (1  $\mu$ g/ml) and  $5 \times 10^4$  APCs in 96-well round-bottom plates for 3 days in the presence of tester cells (CD45.2). The CFSE dilution of CD45.1  $T_{conv}$  cells was assessed by flow cytometry.

### Mammalian luciferase reporter assay

Reporter assays were conducted as previously described<sup>14</sup>. In brief, 50,000 HEK 293 cells per well were plated in 96-well plates in antibiotic-free Dulbecco's Modified Eagle medium (DMEM) containing 1% fetal calf serum (FCS). Cells were transfected with a DNA mixture containing 0.5  $\mu$ g/ml of firefly luciferase reporter plasmid (Promega pGL4.31 (luc2P/Gal4UAS/Hygro)), 2.5 ng/ml of a plasmid containing *Renilla* luciferase (Promega pRL-CMV), and GAL4–DNA binding

# Article

domain–RORY (0.2 µg/ml). Transfections were performed using TransIT-293 (Mirus) according to the manufacturer's instruction. Bile acids or vehicle control were added 24 h after transfection and luciferase activity was measured 16 h later using the dual-luciferase reporter kit (Promega).

## Microscale thermophoresis assay

The binding affinity of the compounds with RORY ligand-binding domain was analysed by microscale thermophoresis (MST). Purified RORY ligand-binding domain was labelled with the Monolith NT Protein Labelling Kit RED (NanoTemper Technologies). Serially diluted compounds, with concentrations of 1 mM to 20 nM, were mixed with 55 nM labelled RORY ligand-binding domain at room temperature and loaded into Monolith standard-treated capillaries. Binding was measured by monitoring the thermophoresis with 20% LED power and 'medium' MST power on a Monolith NT.115 instrument (Nano Temper Technologies) with the following time setting: 5 s Fluo, before; 20 s MST on; and 5 s Fluo, after.  $K_d$  values were fitted using the NT Analysis software (Nano Temper Technologies).

## qPCR with reverse transcription

Total RNA was isolated from cultured T cells using an RNeasy kit (Qiagen) and reverse-transcribed using a PrimeScript RT kit (Takara). All qPCRs were run on the Bio-Rad CFX real-time system using iTaq Universal SYBR Green Supermix (Bio-Rad).  $\beta$ -Actin was used as an internal control to normalize the data across different samples. Primers used for qPCR were as follows: *Foxp3* forward (-F), 5'-ACTGGGGTCTTCTCCCTCAA-3'; *Foxp3* reverse (-R), 5'-CGTGGGAAGGTGCAGAGTAG-3'; *Actb*-F, 5'-CGCC ACCAGTTCGCCATGGA-3'; *Actb*-R, 5'-TACAGCCCGGGAGCATCGT-3'.

## Metabolic assays

In vitro differentiated cells were cultured in the presence of DMSO or isoalloLCA for 48 h, and washed extensively before the assay. The oxygen consumption rate was determined using a Seahorse XF96 Extracellular Flux Analyzer (Seahorse Bioscience) following protocols recommended by the manufacturer and according to the previously published method<sup>52</sup>. In brief, cells were seeded on XF96 microplates (150,000 cells per well) that had been pre-coated with poly-D-lysine (Sigma) to immobilize cells. Cells were maintained in XF medium in a non-CO<sub>2</sub> incubator for 30 min before the assay. The Mito stress test kit (Agilent) was used to test the oxygen consumption rate by sequential injection of 1 µM oligomycin, 1.5 µM FCCP (carbonyl cyanide 4-(trifluoromethoxy)phenylhydrazone) and 0.5 µM rotenone or antimycin A. Data were analysed by wave software (Agilent).

## Chromatin immunoprecipitations

Chromatin immunoprecipitation (ChIP) assays were performed according to a standard protocol. In brief, naive CD4<sup>+</sup> T cells were cultured for 48 h, and fixed for 10 min with 1% formaldehyde. Then, 0.125 M glycine was added to quench the formaldehyde. Cells were lysed, and chromatin was collected and fragmented by sonication at a concentration of 10<sup>7</sup> cells per ChIP sample. Chromatin was immunoprecipitated with 5 µg of ChIP or IgG control antibodies at 4 °C overnight and incubated with protein G magnetic beads (ThermoFisher) at 4 °C for 2 h, washed and eluted in 150 µl elution buffer. Eluate DNA and input DNA were incubated at 65 °C to reverse the crosslinking. After digestion with proteinase K, DNA was purified with the QIAquick PCR purification kit (Qiagen). The relative abundance of precipitated DNA fragments was analysed by qPCR using SYBR Green Supermix (Bio-Rad). The primers used were as follows: *Foxp3* promoter-F, 5'-TAATGTGGCAGTTTCCCAAGCC-3'; *Foxp3* promoter-R, 5'-AATACCTCTCTGCCACTTTCGCCA-3'; *Foxp3* CNS1-F, 5'-AGACTGTCTGGAACAACCTAGCCT-3'; *Foxp3* CNS1-R, 5'-TGGAGGT ACAGAGAGGTTAAGAGCCT-3'; *Foxp3* CNS2-F, 5'-ATCTGGCCAAGTTCA GGTGTGAC-3'; *Foxp3* CNS2-R, 5'-GGGCGTTCCTGTTGACTGTTTCT-3';

*Foxp3* CNS3-F, 5'-TCTCCAGGCTTCAGAGATTCAAGG-3'; *Foxp3* CNS 3-R, 5'-ACAGTGGGATGAGGATACATGGCT-3'; *Foxp3* ex10-F, 5'-CT GCATCGTAGCCACCAGTA-3'; *Foxp3* ex10-R, 5'-AACTATTGCCAT GGCTTCC-3'; *Hsp90ab1*-F, 5'-TTACCTTGACGGGAAAGCCGAGTA-3'; *Hsp90ab1*-R, 5'-TTCGGGAGCTCTCTTGAGTCACC-3'.

## Isolation of lamina propria lymphocytes

Gut tissues were collected and treated with 1 mM DTT at room temperature for 10 min, and 5 mM EDTA at 37 °C for 20 min to remove epithelial cells, and dissociated in digestion buffer (RPMI, 1 mg/ml collagenase type VIII, 100 µg/ml DNase I and 5% FBS) with constant stirring at 37 °C for 30 min. Mononuclear cells were collected at the interface of a 40%–80% Percoll gradient (GE Healthcare). Cells were then analysed by flow cytometry. The distal one-third of the small intestines was considered the ileum.

## Mouse experiments

For bile-acid feeding experiments, the standard mouse diet in ground meal format (PicoLab Diet, no. 5053) was evenly mixed with a measured amount of bile acid compounds and provided in glass feeder jars and replenished when necessary. Bile-acid feeding experiments were performed for seven days or less as extended feeding sometimes resulted in reduced food consumption and weight loss. Colonization of mice with SFB was done with fresh faecal samples, derived from *Il23<sup>-/-</sup> Rag2<sup>-/-</sup>* double-knockout mice that are known to carry much higher levels of SFB compared to conventional B6 mice. Faecal samples were homogenized in water using a 70-µm cell strainer and a 5-ml syringe plunger. Supernatant was introduced into mice using a 20G gavage needle at 250 µl per mouse, approximately equal to the amount of 1/4 mouse faecal pellets. Successful colonization was assessed by qPCR, using the following primers: SFB-F, 5'-GACGCTGAGGCATGAGAG CAT-3'; SFB-R, 5'-GACGGCACGAATTGTTATTCA-3'; universal 16S-F, 5'-ACTCCTACGGGAGGCAGCAGT-3'; universal 16S-R, 5'-ATTACCGCGG CTGCTGGC-3'. For the *C. rodentium* infection experiment, age- and sex-matched germ-free mice were orally infected with approximately 1 × 10<sup>6</sup> colony-forming units of *C. rodentium* and killed for analysis at 6 days after infection. Mice were kept in IsoCage system (Tecniplast) and fed an autoclaved diet with or without 0.3% 3-oxoLCA (w/w) during the experiment.

## Bone marrow transfer

Bone marrow cells were isolated from the femur and tibia of B6 (CD45.1) mice or of CNS3-knockout mice (CD45.2). Red blood cells were removed by using an ammonium-chloride-potassium lysing buffer. The two populations were mixed at a 1:1 ratio and a total of 1 × 10<sup>7</sup> cells were transferred into each irradiated (1,000 rad) CD45.1 mouse (5 weeks old) by retro-orbital injection. Sulfamethoxazole-trimethoprim (240 mg in 250 ml drinking water) was provided for 2 weeks after irradiation.

## Adoptive transfer colitis

CD45RB<sup>high</sup> adoptive transfer colitis was performed as previously described<sup>53</sup>. In brief, isolated CD4<sup>+</sup>CD25<sup>-</sup>CD45RB<sup>high</sup> naive T cells were sorted from wild-type B6 (CD45.1) mice by FACS and 0.5 million cells were adoptively transferred into each RAG1-knockout recipient mouse. In addition, the same number of in vitro cultured and sort-purified CD45.2<sup>+</sup>FOXP3–GFP<sup>+</sup> cells were transferred into the recipient mice. Naive CD4 T cells, isolated from CD45.2 FOXP3–IRES–GFP mice, were cultured under TGFβ<sup>low</sup> (0.05 ng/ml TGFβ), isoalloLCA (20 µM isoalloLCA and 0.01 ng/ml TGFβ) or TGFβ<sup>high</sup> (1 ng/ml TGFβ) conditions. Mice were then monitored and weighed each week. At week 8, colon tissues were collected, and lamina propria lymphocytes were analysed by flow cytometry. Haematoxylin and eosin (H & E) staining and disease scoring were performed by the Rodent Histopathology Core at Harvard Medical School.

## Isolation of faecal bacterial microbiota and 16S rRNA gene sequencing analysis

DNA of the mouse faecal microbiota was isolated by using QIAamp Fast DNA Stool Mini Kit (Qiagen) according to the manufacturer's instructions. The samples were quantified using an Agilent 4200 TapeStation instrument, with corresponding Agilent Genomic DNA ScreenTape assays. The samples were then normalized to 12.5 ng of input in 2.5 µl (5 ng/µl), and amplified using IDT primers specific to the V3 and V4 region: forward 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTACGGGNGGCWGCAG-3', reverse 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC-3'. The amplification was done using the KAPA HiFi HotStart Ready Mix (2×) (Roche Sequencing Solutions). Residual primers were eluted away using Aline PCR Clean DX beads in a 0.8× SPRI-based cleanup. The purified amplicons were then ligated with indexing adapters using Illumina's Nextera XT Index Primers. Following this step, a final clean-up was performed using Aline PCR Clean DX beads. The resulting purified libraries were run on an Agilent 4200 TapeStation instrument, with a corresponding Agilent High Sensitivity D1000 ScreenTape assay to visualize the libraries and check that the size of the library matched the expected about 630-bp product. Concentrations obtained from this assay were used to normalize all samples in equimolar ratio. The pool was denatured and loaded onto an Illumina MiSeq instrument, with an Illumina MiSeq V3 600-cycle kit to obtain paired-end 300-bp reads. The pool was loaded at 10.5 pM, with 50% PhiX spiked in to compensate for low base diversity. The basecall files were demultiplexed through the BioPolymer Facility's pipeline, and the resulting FASTQ files were used in subsequent analysis. Raw fastq sequences were then quality-filtered and analysed by following QIIME2 version 2018.11 and DADA2 1.6.0<sup>54–56</sup>. Operational taxonomic units were picked with 97% sequence similarity. The phylogenetic affiliation of each operational taxonomic unit was aligned to the Greengenes reference database version 13.8 and 99% ID.

## Statistical analyses

Statistical analysis tests were performed with Prism v.8.0.2 (GraphPad).

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

The 16S rDNA datasets are available through NCBI under accession number PRJNA528994. Source Data for Figs. 1–4 and Extended Data Figs. 2–9 are provided with the paper. Any other relevant data are available from the corresponding authors upon reasonable request.

51. Yao, L. et al. A selective gut bacterial bile salt hydrolase alters host metabolism. *eLife* **7**, e37182 (2018).
52. van der Windt, G. J., Chang, C. H. & Pearce, E. L. Measuring bioenergetics in T cells using a Seahorse extracellular flux analyzer. *Curr. Protoc. Immunol.* **113**, 3.16B.1–13.16B.14 (2016).
53. Powrie, F. et al. Inhibition of Th1 responses prevents inflammatory bowel disease in scid mice reconstituted with CD45RB<sup>hi</sup> CD4<sup>+</sup> T cells. *Immunity* **1**, 553–562 (1994).
54. Bokulich, N. A. et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* **6**, 90 (2018).
55. Bolyen, E. et al. An introduction to applied bioinformatics: a free, open, and interactive text. *J Open Source Educ* **1**, 27 (2018).
56. Callahan, B. J. et al. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).

**Acknowledgements** We thank N. Lee and K. Hattori for technical assistance; M. Trombly for critical reading of the manuscript; A. Rudensky and S. Smale for sharing FOXP3–CNS- and REL- knockout mice; R. Bronson and the Rodent Histopathology Core at Harvard Medical School for performing H & E analysis and disease score; the BPF Next-Gen Sequencing Core at Harvard Medical School for their expertise and instrument availability with microbiota sequencing. We acknowledge NIH grant P30DK034854 and the use of the Harvard Digestive Disease Center's (HDDC's) core services, resources, technology and expertise. This study was supported by a Charles A. King Trust Fellowship to S. Hang, Harvard Medical School Dean's Innovation Grant in the Basic and Social Sciences to A.S.D. and J.R.H., the Howard Hughes Medical Institute to D.R.L. and National Institutes of Health grants R01AI080885 to D.R.L. and R01DK110559 to J.R.H.

**Author contributions** M.A.F., J.R.H., and D.R.L. conceptualized the study. S. Hang, D.P., A.S.D., M.R.K., M.A.F., D.R.L., and J.R.H. conceived and designed the experiments; S. Hang and D.P. performed most of the experiments; L.Y., E.K., T.J., A.S.D., J.L., S. Ha, B.N.N., S.P.K., and L.W. provided help with experiments; J.L. and F.R. designed and performed the RORyt binding assay; B.N.N., S.P.K., and M.R.K. synthesized some of the bile acid derivatives; L.Y. and A.S.D. performed in vivo bile acid analyses; R.S.L. and Y.Z. provided critical materials; and S. Hang, D.P., and J.R.H. wrote the manuscript, with contributions from all authors.

**Competing interests** A.S.D. is an ad hoc consultant for Kintai Therapeutics. D.R.L. is a scientific co-founder of Vedanta Biosciences.

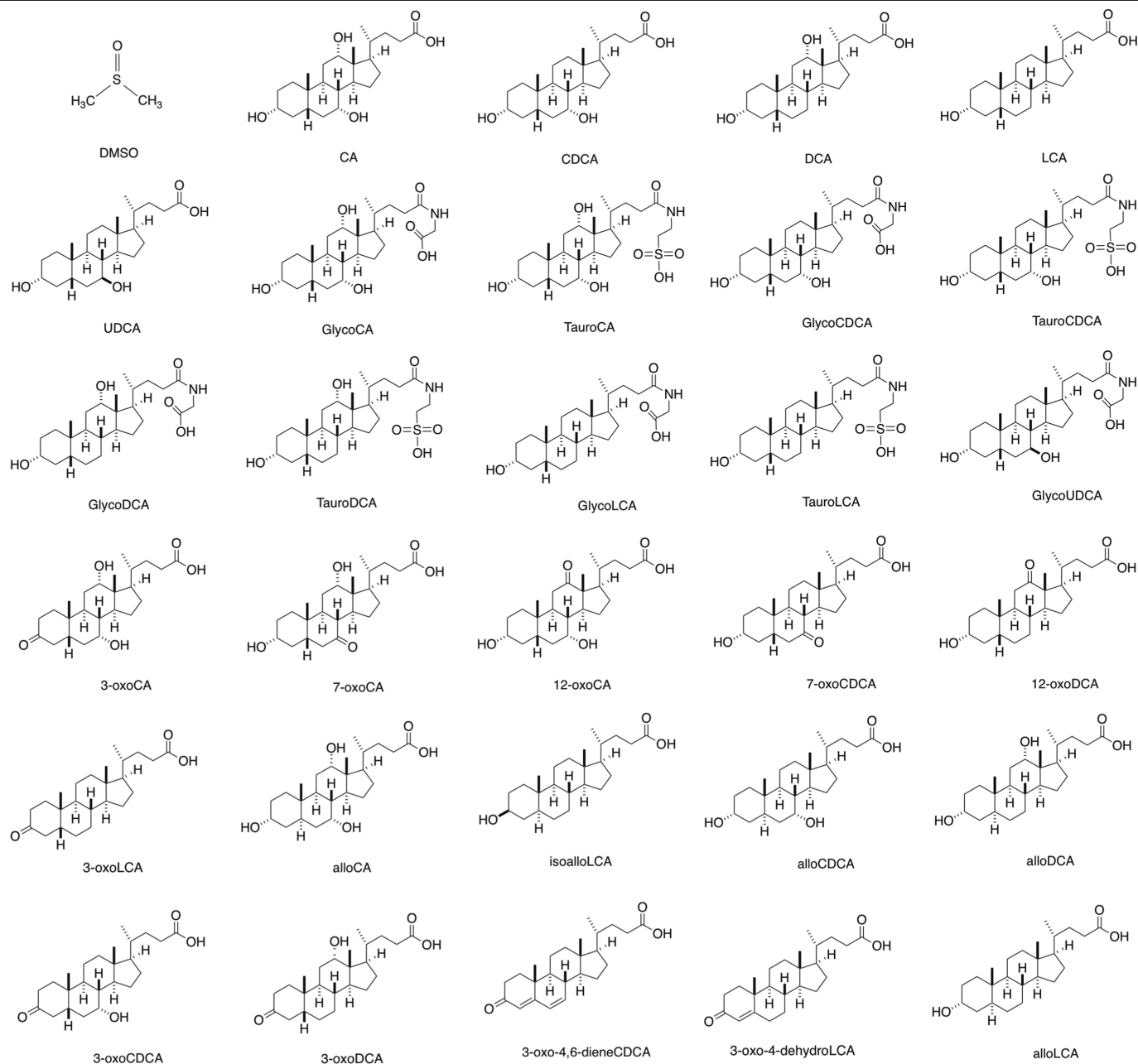
## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1785-z>.

**Correspondence and requests for materials** should be addressed to M.A.F., D.R.L. or J.R.H.

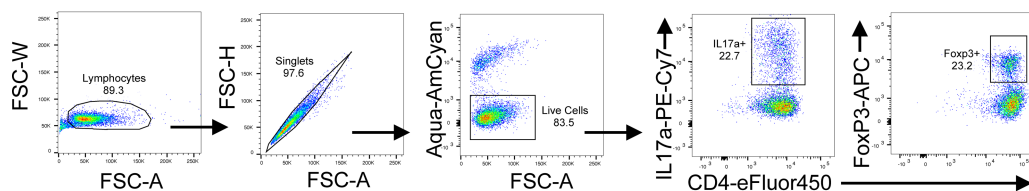
**Peer review information** Nature thanks Navdeep S. Chandel, Richard Steven Blumberg and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

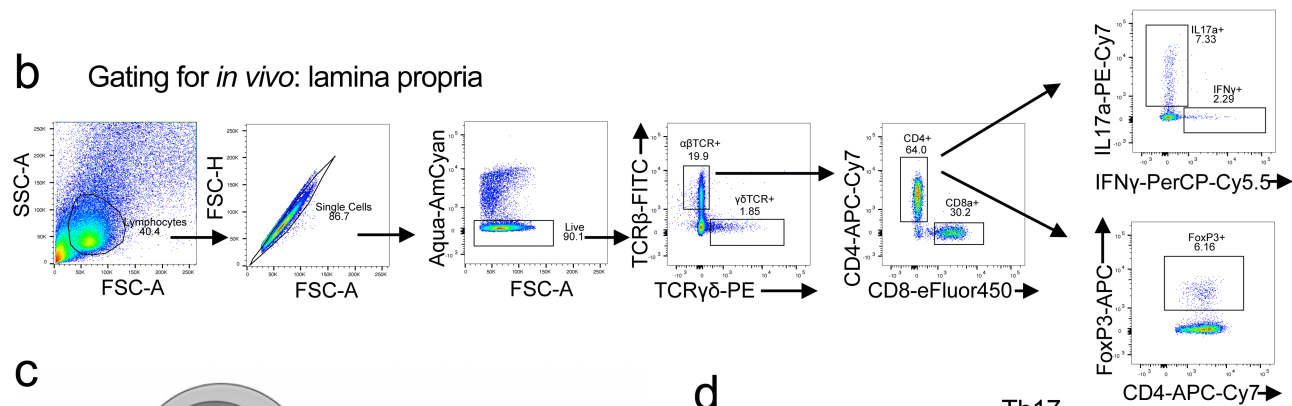


**Extended Data Fig. 1 | Chemical structures of bile acid derivatives.** These derivatives were used for the T cell differentiation assay.

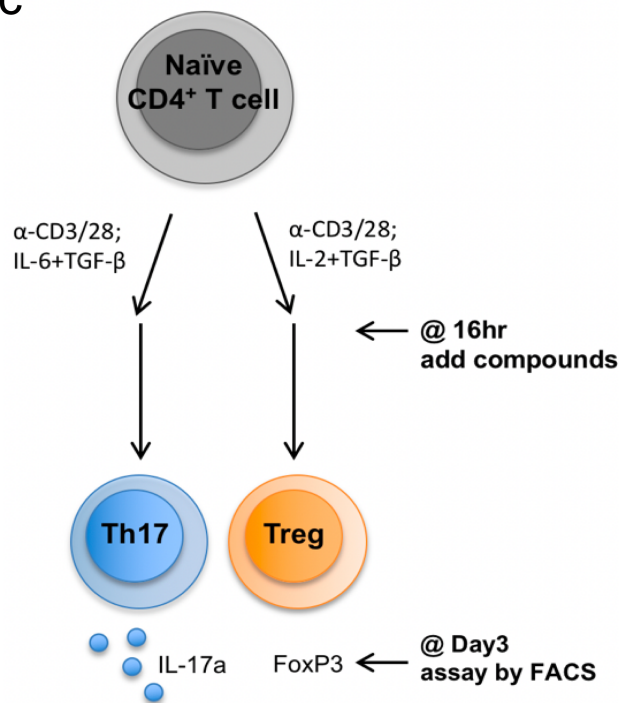
## a Gating for *in vitro* culture



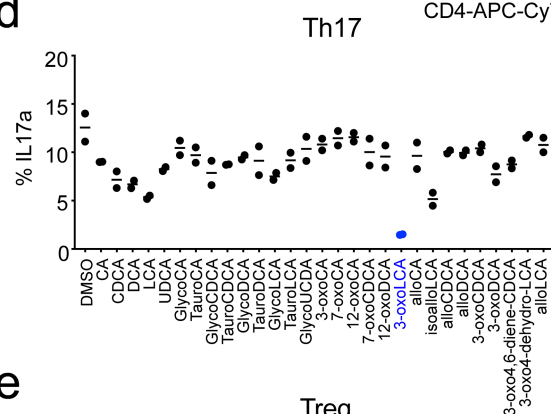
## b Gating for *in vivo*: lamina propria



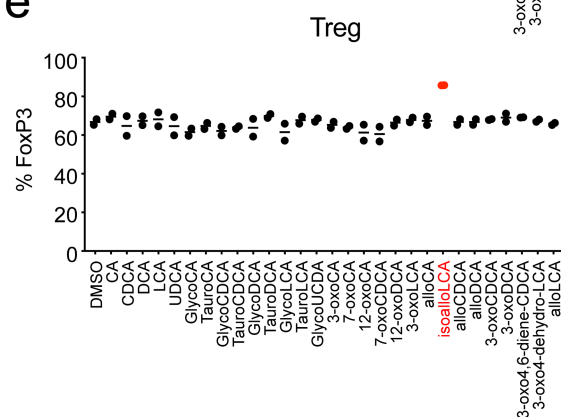
## c



## d



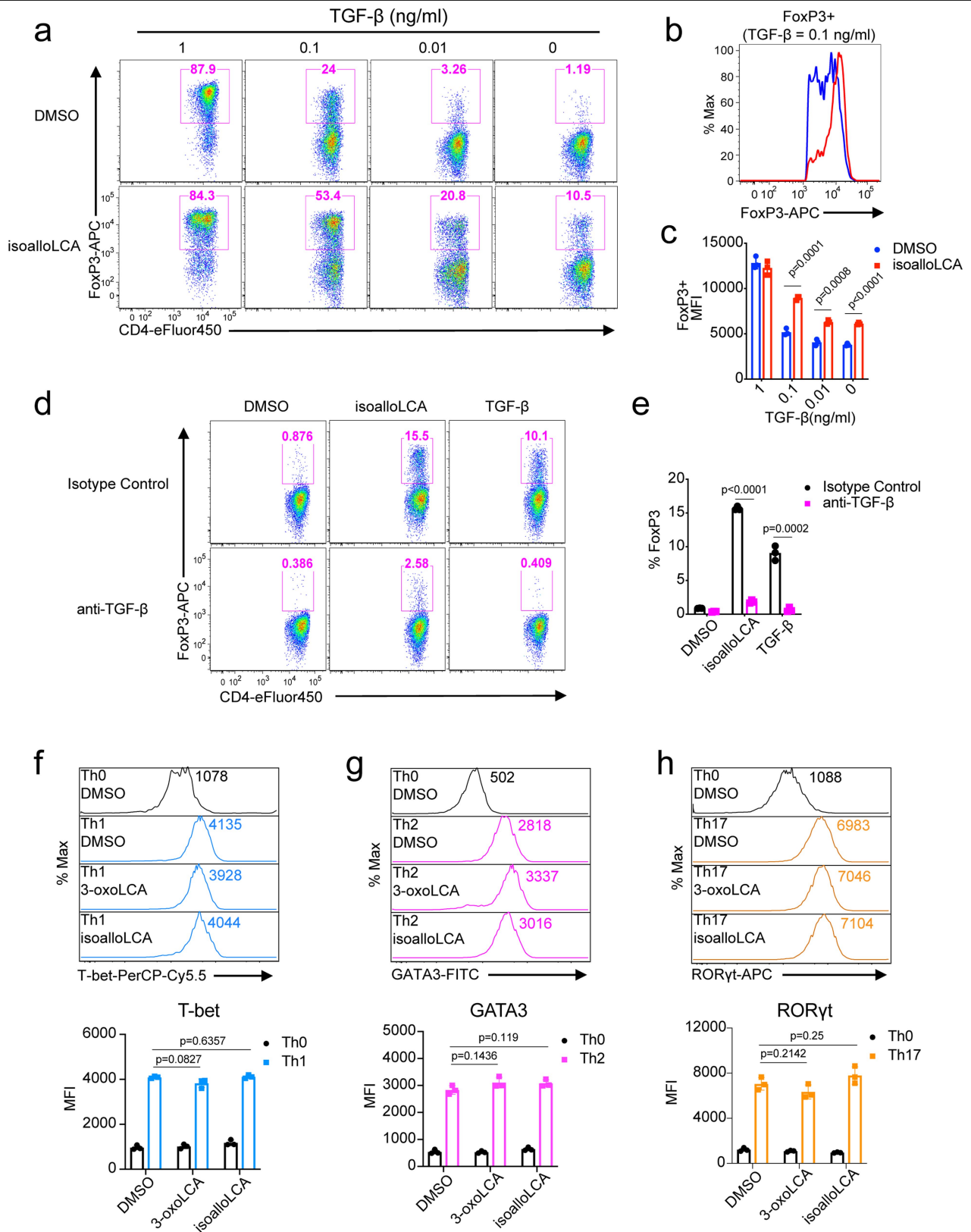
## e



**Extended Data Fig. 2 | 3-OxoLCA and isoalloLCA affect  $T_H17$  and  $T_{reg}$  cell differentiation.** **a, b**, Gating strategy for the flow cytometric analyses of *in vitro* cultured T cells (**a**) and *in vivo* derived cells from the lamina propria (**b**). **c**, Schematic of the screening procedure. **d, e**, Naïve  $CD4^+$  T cells isolated from B6 Jax mice ( $n = 2$  biologically independent samples) were cultured under  $T_H17$

(IL-6 = 10 ng ml<sup>-1</sup>; TGFβ = 0.5 ng ml<sup>-1</sup>) (**d**) and  $T_{reg}$  (IL-2 = 100 U ml<sup>-1</sup>; TGFβ = 0.1 ng ml<sup>-1</sup>) (**e**) cell polarization conditions for 3 days. DMSO or various bile acids at 20 μM concentration were added to the cell cultures on day 1. Data are mean.

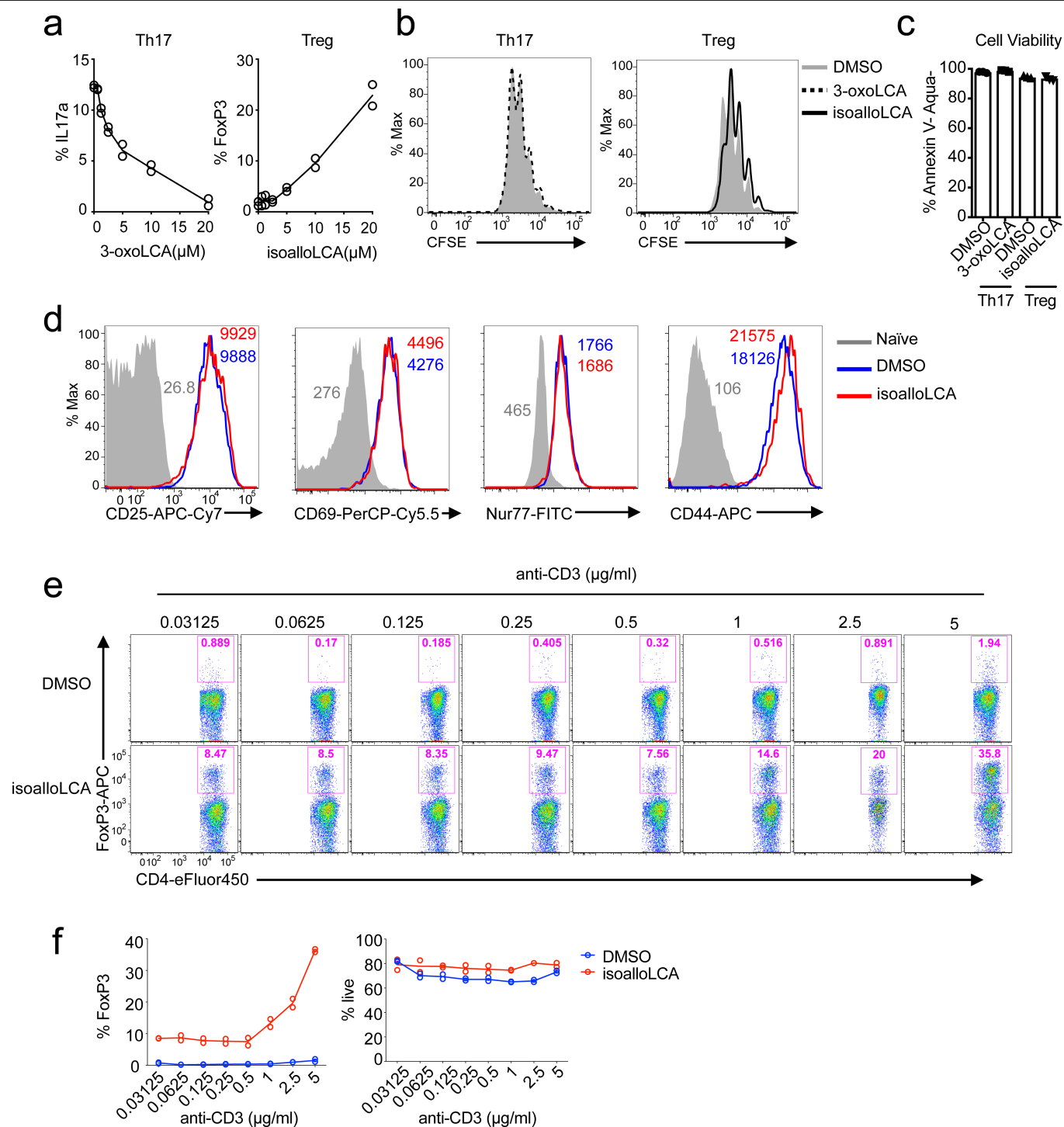




**Extended Data Fig. 3 | IsoalloLCA-induced T<sub>reg</sub> cell expansion requires TGFβ.**

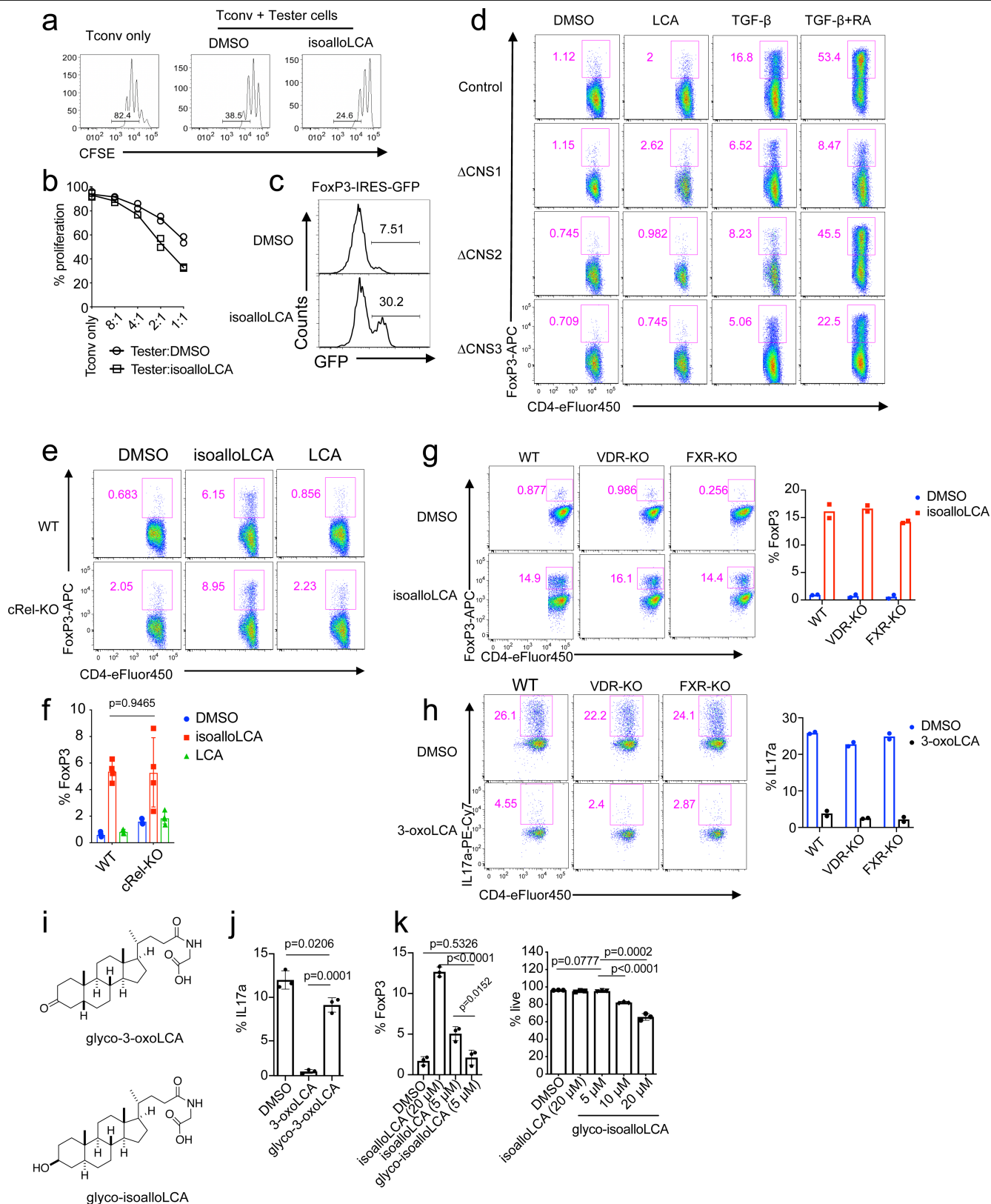
**a–c**, Flow cytometry and histogram of CD4<sup>+</sup> T cells, cultured for 3 days with different amounts of TGFβ (1, 0.1, 0.01 or 0 ng ml<sup>-1</sup>) and IL-2 (100 U ml<sup>-1</sup>) in the presence of DMSO or isoalloLCA (20 μM) and intracellularly stained for FOXP3 (*n* = 3 biologically independent samples per group). **d, e**, Flow cytometry of CD4<sup>+</sup> T cells, cultured for 3 days in the presence of DMSO, isoalloLCA (20 μM) or TGFβ (0.05 ng ml<sup>-1</sup>). In addition, anti-TGFβ antibody (10 μg ml<sup>-1</sup>, 1D11) or isotype control were added to the culture (*n* = 3 biologically independent samples per

group). **f–h**, 3-OxoLCA and isoalloLCA do not affect key transcription factor expression. T cells were cultured under T<sub>H</sub>0, T<sub>H</sub>1, T<sub>H</sub>2 or T<sub>H</sub>17 conditions, in the presence of DMSO, 3-oxoLCA (20 μM) or isoalloLCA (20 μM). T-cell-lineage-determining transcription factors such as T-bet, GATA3 or RORγt were intracellularly stained (*n* = 3 biologically independent samples per group). MFI, mean fluorescence intensity. Data are mean ± s.d., by unpaired *t*-test with two-tailed *P* value.



**Extended Data Fig. 4 | Effects of isoalloLCA on FOXP3 expression require strong TCR stimulation.** **a**, 3-OxoLCA and isoalloLCA demonstrate dose-dependent effects on Th17 cell and Treg cell differentiation, respectively ( $n = 2$  biologically independent samples). A low concentration of TGF $\beta$  ( $0.01 \text{ ng ml}^{-1}$ ) was used for Treg cell culture. **b–d**, 3-OxoLCA and isoalloLCA do not significantly affect cell proliferation, cell viability or T cell activation. **b**, Naïve CD4<sup>+</sup> T cells were labelled with a cell proliferation dye CFSE and cultured for 3 days in the presence of DMSO, 3-oxoLCA or isoalloLCA under Th17 or Treg cell polarization conditions. **c**, Live-cell percentages at the end of the 3-day culture were determined based on both annexin V and fixable live/

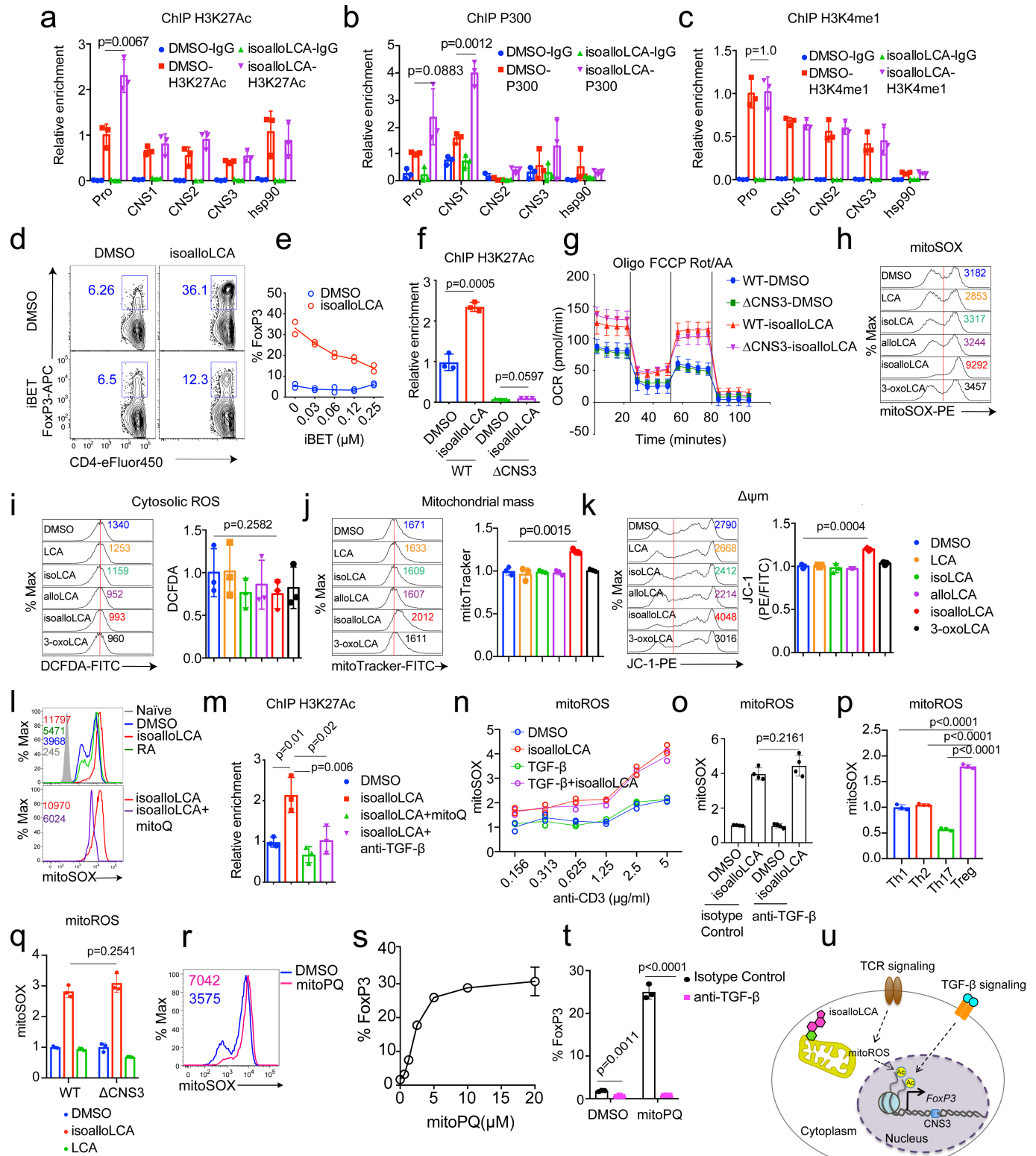
dead staining ( $n = 3$  biologically independent samples per group). **d**, Both DMSO and isoalloLCA treatment lead to comparable levels of expression of CD25, CD69, NUR77 and CD44. Naïve CD4<sup>+</sup> T cells were used as a negative control. **e**, **f**, T cells were cultured with different concentrations of anti-CD3 antibody, in the presence of DMSO or isoalloLCA (20  $\mu\text{M}$ ). Representative FACS plots of CD4<sup>+</sup> T cells cultured for 3 days and stained intracellularly for FOXP3 (**e**). Quantification of FOXP3<sup>+</sup> and viable T cells after 3-day culture (**f**) ( $n = 2$  biologically independent samples per group). Data are representative of two independent experiments (**b**, **d**). Data in **c** are mean  $\pm$  s.d.



Extended Data Fig. 5 | See next page for caption.

**Extended Data Fig. 5 | REL, VDR and FXR are dispensable for isoalloLCA-dependent induction of FOXP3.** **a, b**, In vitro suppression assay. CD4<sup>+</sup> effector T cells (T<sub>conv</sub>) were labelled with CFSE and mixed with DMSO- or isoalloLCA-treated T<sub>reg</sub> cells (tester) at different ratios (*n* = 2 biologically independent samples per group). **c**, Expression of GFP in DMSO- or isoalloLCA-treated T cells cultured with anti-CD3/28, IL-2 and TGFβ (0.01 ng ml<sup>-1</sup>). Naive CD4<sup>+</sup> T cells were isolated from FOXP3-IRES-GFP mice. **d**, Flow cytometry of CD4<sup>+</sup> T cells stained intracellularly for FOXP3. Naive CD4<sup>+</sup> T cells isolated from wild-type, CNS1-, CNS2- or CNS3-knockout mice (*n* = 3 biologically independent samples per group) were cultured with anti-CD3/28 and IL-2, LCA (20 μM), TGFβ (0.05 ng ml<sup>-1</sup>) and additional retinoic acid (1 ng ml<sup>-1</sup>). **e, f**, Flow cytometry (**e**) and its quantification (**f**) of CD4<sup>+</sup> T cells stained intracellularly for FOXP3. Naive CD4<sup>+</sup> T cells were isolated from wild-type control mice or REL-knockout mice (*n* = 4 biologically independent samples per group) and cultured with anti-CD3/28 and IL-2 in the presence of DMSO, isoalloLCA (20 μM) or LCA (20 μM).

**g, h**, Naive CD4<sup>+</sup> T cells isolated from wild-type control, VDR-knockout or FXR-knockout (*n* = 2 biologically independent samples per group) were cultured with anti-CD3/28 and IL-2 (**g**) or anti-CD3/28, IL-6 and TGFβ (**h**) for 3 days in the presence of DMSO, isoalloLCA (20 μM), or 3-oxoLCA (20 μM). Representative FACS plots of T cells intracellularly stained for FOXP3 or IL-17a. **i**, Chemical structures of glycine-conjugated 3-oxoLCA (glyco-3-oxoLCA) and isoalloLCA (glyco-isoalloLCA). **j** and **k**, Quantifications of T<sub>H</sub>17 (**j**) and T<sub>reg</sub> (**k**) cell differentiation in vitro. T cells were cultured with anti-CD3/28, IL-6 and TGFβ (**j**) or anti-CD3/28 and IL-2 (**k**) in the presence of DMSO, 3-oxoLCA (20 μM), glyco-3-oxoLCA (20 μM), isoalloLCA (5 or 20 μM) or glyco-isoalloLCA (5, 10 or 20 μM). Glyco-isoalloLCA exhibited enhanced cytotoxicity at 10 or 20 μM compared to isoalloLCA (*n* = 3 biologically independent samples per group). Data are representative of two independent experiments (**c, d**). Data are mean ± s.d., by unpaired *t*-test with two-tailed *P* value.

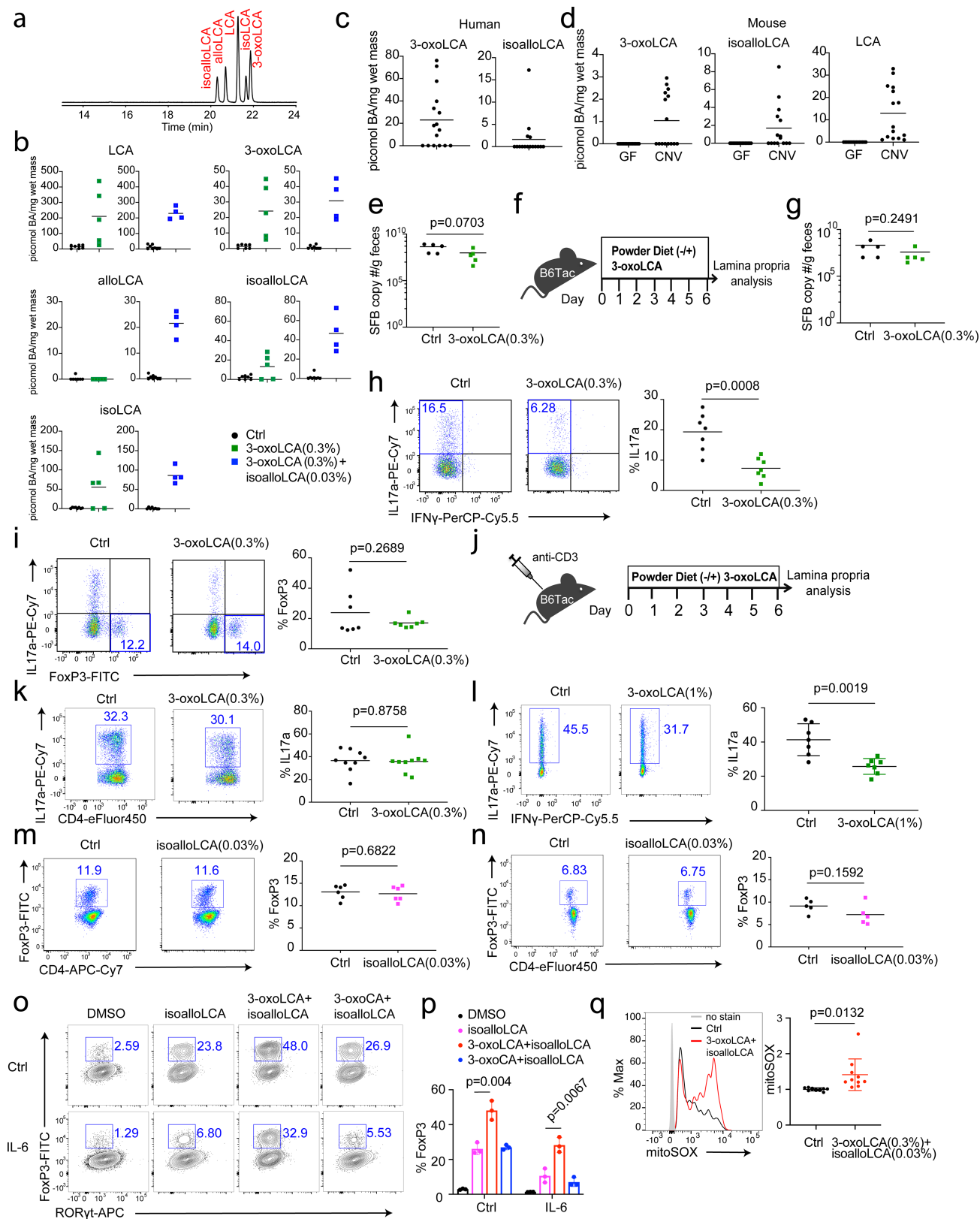


Extended Data Fig. 6 | See next page for caption.



**Extended Data Fig. 6 | IsoalloLCA-dependent FOXP3 transcription requires mitoROS and H3K27ac.** **a–c**, ChIP analysis of H3K27ac, p300 and H3K4 mono-methylation (H3K4me1) on the *Foxp3* gene locus. Chromatin obtained from DMSO- and isoalloLCA-treated wild-type cells were immunoprecipitated with IgG, anti-H3K27ac, anti-p300 or anti-H3K4me1 antibodies, followed by real-time PCR analysis ( $n = 3$  biologically independent samples per group). Primers targeting *Foxp3* promoter (Pro), CNS1, CNS2 and CNS3 region and *Hsp90ab1* promoter were used for qPCR quantification. Relative enrichment was calculated as fold change relative to the ChIP signal at the *Foxp3* promoter of the DMSO-treated control. **d, e**, Flow cytometry and quantification of CD4<sup>+</sup> T cells stained intracellularly for FOXP3. Naive CD4<sup>+</sup> T cells isolated from wild-type mice ( $n = 2$  biologically independent samples per group) were cultured with anti-CD3/28, IL-2 and TGF $\beta$  (0.05 ng ml<sup>-1</sup>) in the presence of DMSO or isoalloLCA (20  $\mu$ M) in the presence or absence of iBET. **f**, ChIP analysis of H3K27ac on the *Foxp3* promoter region. Naive CD4<sup>+</sup> T cells isolated from wild-type or CNS3-knockout mice ( $n = 3$  biologically independent samples per group) were treated with DMSO or isoalloLCA (20  $\mu$ M). **g**, Seahorse analysis of oxygen consumption rate (OCR) with naive CD4<sup>+</sup> T cells isolated from wild-type or CNS3-knockout mice cultured with anti-CD3/28 and IL-2 for 48 h, in the presence of DMSO or isoalloLCA (20  $\mu$ M). Measurements from six wells from two mice for each genotype. **h–k**, T cells were cultured with DMSO, LCA, isoLCA, alloLCA, isoalloLCA or 3-oxoLCA at 20  $\mu$ M for 48 h. Their mitochondrial and cytoplasmic ROS were measured by mitoSOX (**h**) and 2',7'-dichlorofluorescein diacetate (DCFDA) (**i**), respectively. Total mitochondria mass was measured by MitoTracker (**j**) and the mitochondrial membrane potential measured by JC-1 dye (**k**). Mean fluorescence intensities of

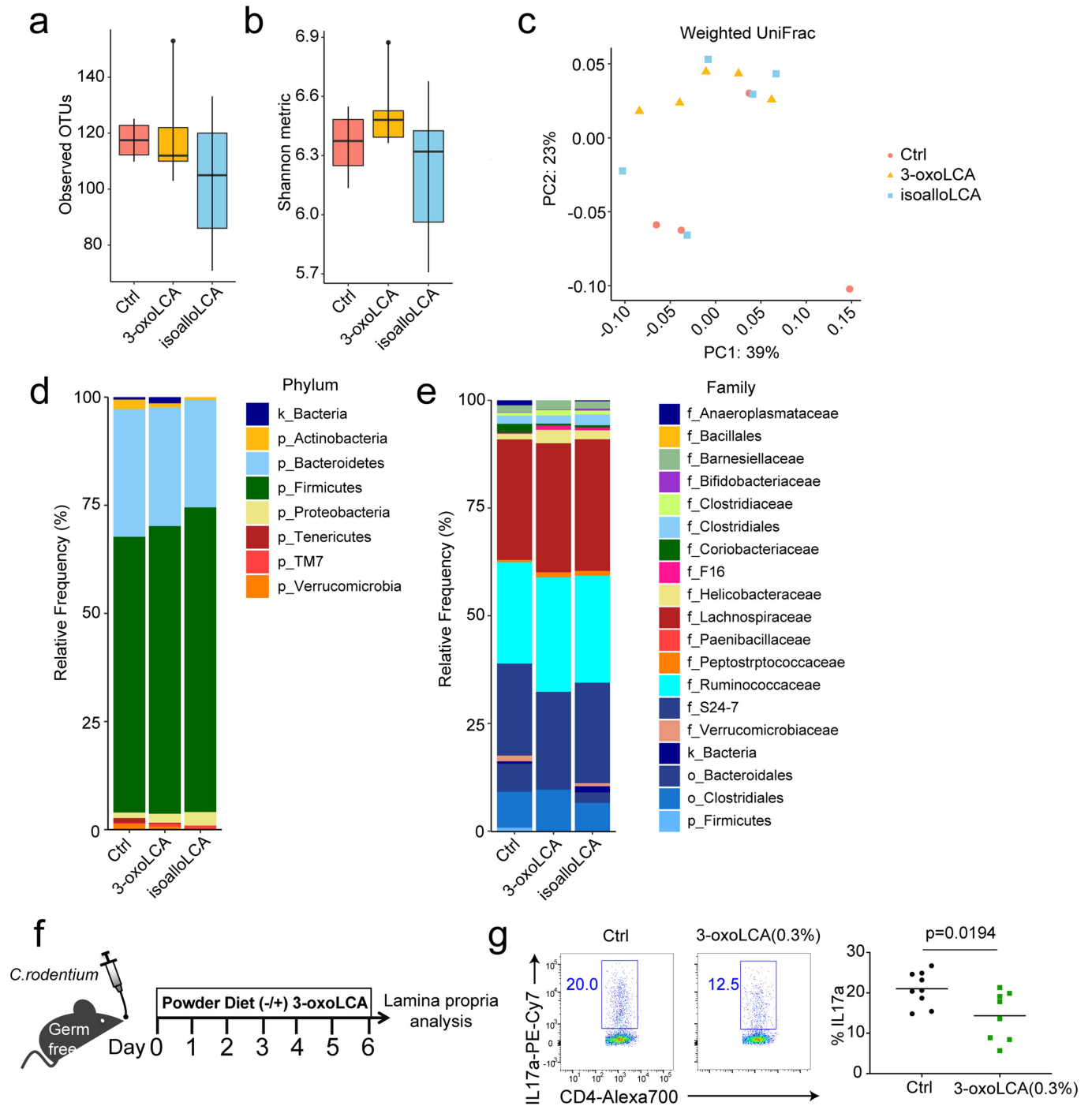
different treatments were normalized as fold changes of those of the DMSO control ( $n = 3$  biologically independent samples per group). **l**, MitoROS production measured by mitoSOX with T cells cultured with DMSO, isoalloLCA (20  $\mu$ M), retinoic acid (1 nM), or isoalloLCA (20  $\mu$ M) + mitoQ (0.5  $\mu$ M) for 48 h. **m**, ChIP analysis ( $n = 3$  biologically independent samples per group) of H3K27ac on the *Foxp3* promoter of T cells, treated with DMSO, isoalloLCA, isoalloLCA + mitoQ or isoalloLCA + anti-TGF $\beta$  for 72 h. **n–q**, MitoROS production measured by mitoSOX with T cells cultured with different concentrations of anti-CD3 and treated with DMSO, isoalloLCA (20  $\mu$ M), TGF $\beta$  (0.05 ng ml<sup>-1</sup>) or isoalloLCA plus TGF $\beta$  ( $n = 2$  biologically independent samples per group) (**n**); or with T cells treated with DMSO or isoalloLCA (20  $\mu$ M) plus an isotype control or anti-TGF $\beta$  antibody ( $n = 4$  biologically independent samples per group) (**o**); or with T cells cultured under T<sub>H</sub>1, T<sub>H</sub>2, T<sub>H</sub>17 or T<sub>reg</sub> cell conditions ( $n = 3$  biologically independent samples per group) (**p**); or with naive CD4<sup>+</sup> T cells isolated from wild-type or CNS3-knockout mice and cultured with anti-CD3/28 and IL-2 ( $n = 3$  biologically independent samples per group) (**q**). **r**, MitoROS production measured by mitoSOX with T cells cultured with DMSO or mitoPQ (5  $\mu$ M) for 48 h. **s**, Dose-dependent effects of mitoPQ on T<sub>reg</sub> cell differentiation ( $n = 3$  biologically independent samples per group). **t**, Quantification of T<sub>reg</sub> cell differentiation in vitro on naive CD4<sup>+</sup> T cells cultured in the presence of DMSO or mitoPQ (5  $\mu$ M) and treated with isotype control or anti-TGF $\beta$  antibody ( $n = 3$  biologically independent samples per group). **u**, A model showing the mechanism of isoalloLCA enhancement of T<sub>reg</sub> cell differentiation. Data are representative of two independent experiments (**l, r**) and shown as mean  $\pm$  s.d., by unpaired *t*-test with two-tailed *P* value.



Extended Data Fig. 7 | See next page for caption.

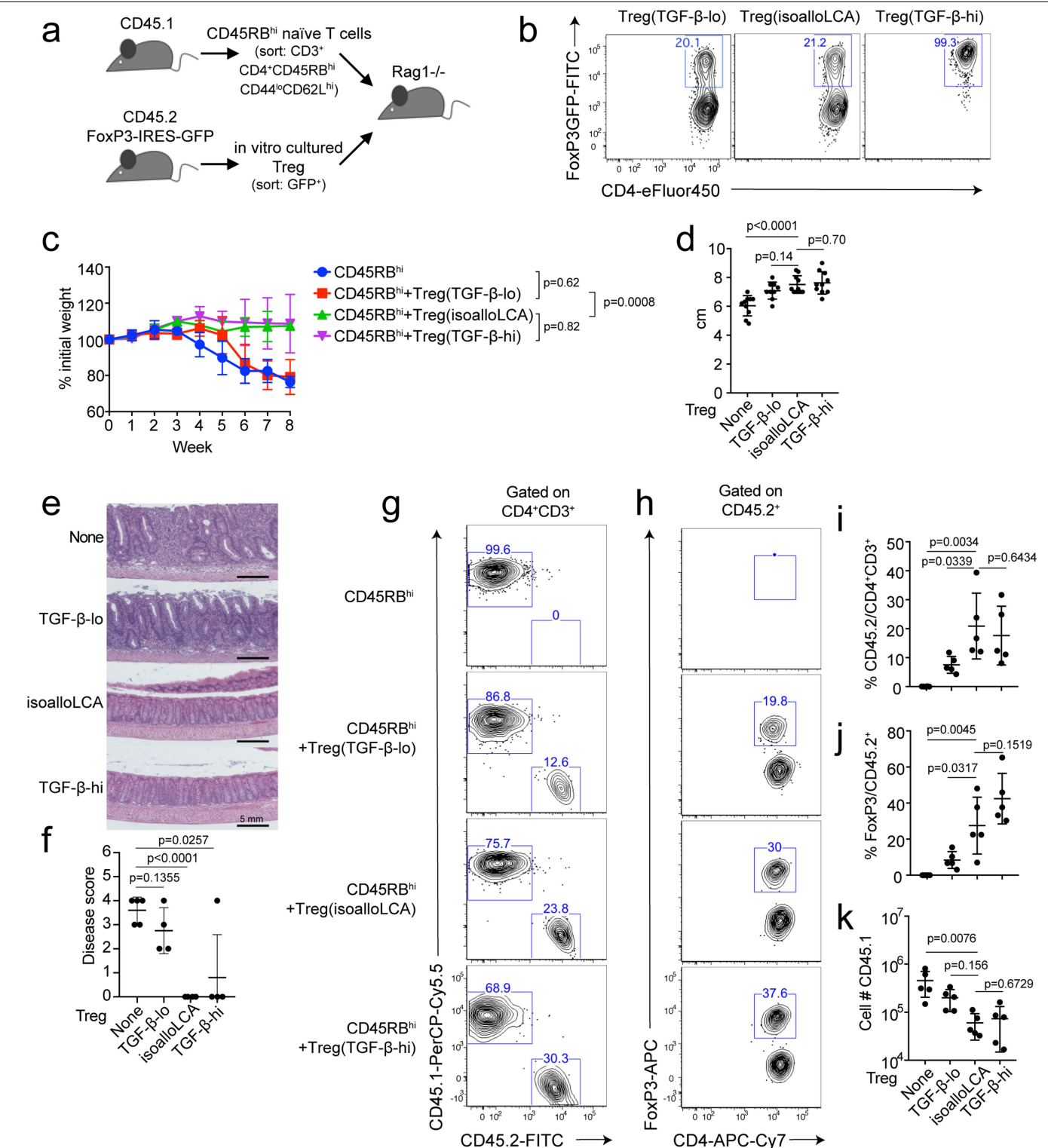
**Extended Data Fig. 7 | 3-OxoLCA inhibits the differentiation of T<sub>H</sub>17 cells but not T<sub>reg</sub> cells, and isoalloLCA alone does not enhance T<sub>reg</sub> cell differentiation in vivo.** **a**, UPLC–MS spectra of LCA and its isomers isoalloLCA, alloLCA, and isoLCA, as well as 3-oxoLCA. **b**, Quantification of unconjugated LCA and its derivatives in the caecal contents of B6 Tac mice fed on a control or bile-acid-containing diet ( $n = 7, 5$  and  $4$  mice for control (ctrl), 3-oxoLCA and 3-oxoLCA + isoalloLCA, respectively). **c**, Quantification of unconjugated 3-oxoLCA and isoalloLCA in human stool samples from patients with ulcerative colitis ( $n = 16$  donors). **d**, Quantification of unconjugated 3-oxoLCA, isoalloLCA and LCA in mouse caecal contents from germ-free (GF) or conventionally housed (CNV) mice ( $n = 15$  mice per group). **e**, B6 Jax mice gavaged with SFB. SFB colonization measured by qPCR analysis calculated as copy number ( $n = 5$  mice per group). **f**, Diagram showing experimental design. B6 Tac mice were fed a 3-oxoLCA (0.3%)-containing diet for 7 days. **g**, SFB colonization measured by qPCR analysis calculated as SFB copy number ( $n = 5$  mice per group). **h, i**, Flow cytometric analysis and quantification of T<sub>H</sub>17 (**h**) and T<sub>reg</sub> (**i**) cells of the ileal lamina propria ( $n = 7$  mice per group). **j–l**, Experimental scheme of anti-CD3 experiment with 3-oxoLCA (**j**). Flow cytometric analysis and quantification of

CD4<sup>+</sup> cells of the lamina propria following an anti-CD3 injection from B6 Tac mice fed with control or 3-oxoLCA (0.3%) diet ( $n = 9$  mice per group) (**k**), or 3-oxoLCA (1%) diet ( $n = 7$  mice per group) (**l**). **m, n**, Flow cytometric analysis and quantification of CD4<sup>+</sup> cells of the ileal lamina propria in steady-state (**m**) ( $n = 6$  mice per group) or following an anti-CD3 injection (**n**) ( $n = 5$  mice per group). B6 Tac mice were fed with control or isoalloLCA (0.03%) diet. **o, p**, Flow cytometry (**o**) and quantification (**p**) of CD4<sup>+</sup> T cells stained intracellularly for FOXP3, showing that the combination of 3-oxoLCA and isoalloLCA further increases T<sub>reg</sub> cell differentiation. Naive CD4<sup>+</sup> T cells isolated from wild-type B6 mice ( $n = 3$  biologically independent samples) treated with DMSO, isoalloLCA (20  $\mu$ M), a mixture of 3-oxoLCA (20  $\mu$ M) and isoalloLCA (20  $\mu$ M) or a mixture of 3-oxoLCA (20  $\mu$ M) and isoalloLCA (20  $\mu$ M) and cultured with anti-CD3/28 and IL-2, with or without the addition of IL-6 (62.5 pg ml<sup>-1</sup>). **q**, MitoROS production in total CD4<sup>+</sup> T cells isolated from the ileal lamina propria. Mice were fed a control diet or diet containing a mixture of 3-oxoLCA (0.3%) + isoalloLCA (0.03%) ( $n = 9$  or  $10$  mice, respectively) and injected with 10  $\mu$ g of anti-CD3 to induce inflammation. Data are mean  $\pm$  s.d., by unpaired  $t$ -test with two-tailed  $P$  value.



**Extended Data Fig. 8 | 3-OxoLCA or isoalloLCA does not significantly alter gut microbiota.** **a**, Box plot showing operational taxonomic unit (OTU) numbers. **b**, Shannon diversity of faecal microbiota based on 16S rRNA gene amplicon sequencing. For the box plots in **a**, **b**, the three horizontal lines of the box represent the third quartile, median and first quartile, respectively, from top to bottom. The whiskers above and below the box show the maximum and minimum. **c**, Principal coordinates analysis based on weighted UniFrac distances of 16S rRNA amplicon sequencing of faecal microbiota. **d**, **e**, Average

relative abundance of microbiota at the phylum (**d**) and the family (**e**) levels by taxon-based analyses ( $n = 4, 5$  and  $5$  mice for the control, 3-oxoLCA and isoalloLCA groups, respectively). **f**, **g**, Experimental scheme (**f**) and flow cytometric analysis and quantification (**g**) of CD4<sup>+</sup> cells of the lamina propria of the colon in germ-free B6 mice, infected with *C. rodentium*. Mice were fed an autoclaved diet with or without 3-oxoLCA (0.3%) ( $n = 9$  mice per group). Data are mean  $\pm$  s.d., by unpaired *t*-test with two-tailed *P* value.



**Extended Data Fig. 9 | IsoalloLCA-induced T<sub>reg</sub> cells suppress transfer colitis.**

**a**, Experimental scheme. Rag1<sup>-/-</sup> recipient mice were transferred intraperitoneally with 0.5 million CD45RB<sup>hi</sup> naïve CD4<sup>+</sup> T cells (CD45.1) and with or without co-transfer of 0.5 million FOXP3-GFP<sup>+</sup> T<sub>reg</sub> cells (CD45.2). FOXP3-GFP<sup>+</sup> cells were cultured under TGF $\beta$ <sup>low</sup> (0.05 ng ml<sup>-1</sup>), isoalloLCA (20  $\mu$ M, 0.01 ng ml<sup>-1</sup> TGF $\beta$ ) and TGF $\beta$ <sup>high</sup> (1 ng ml<sup>-1</sup>) conditions with GFP<sup>+</sup> naïve CD4 T cells, isolated from CD45.2 FOXP3-IRES-GFP mice. **b**, Flow cytometric analysis of the FOXP3-GFP<sup>+</sup> cells, following in vitro culture. The gated cells were sorted and used for co-transfer. **c-f**, Weight change monitored for

8 weeks; week-7 values are used for unpaired *t*-test with two-tailed *P* value (**c**) (*n* = 5 mice per group). At the end of the experiment, colon length (**d**) (*n* = 10 mice per group), H & E staining (**e**) and the quantification of disease score (**f**) (*n* = 5 mice for 'none', 4 mice for other groups). **g-j**, Flow cytometric analysis and quantification of the frequency of CD45.1 and CD45.2 (**g, i**) and the frequency of FOXP3<sup>+</sup> cells in the CD45.2 population (**h, j**) in each condition (*n* = 5 mice per group). **k**, Quantification of total CD45.1 cell number in the lamina propria of the colon (*n* = 5 mice per group). Data are mean  $\pm$  s.d., by unpaired *t*-test with two-tailed *P* value.



Extended Data Table 1 | Lipophilicity of bile acids

bile acid	abbreviation	log D (pH = 8.0)	reference
lithocholic acid	LCA	3.6	This study
isolithocholic acid	isoLCA	3.5	This study
allolithocholic acid	alloLCA	3.5	This study
isoallolithocholic acid	isoalloLCA	2.2	This study
3-oxolithocholic acid	3-oxoLCA	2.4	This study
deoxycholic acid	DCA	2.7	20
chenodeoxycholic acid	CDCA	2.2	20,21
ursodeoxycholic acid	UDCA	2.2	20,21
obeticholic acid	OCA	2.5	21
cholic acid	CA	1.1	20,21

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- |                                     |  |
|-------------------------------------|--|
| n/a                                 | Confirmed  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of all covariates tested  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. $F$ , $t$ , $r$ ) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted<br><i>Give <math>P</math> values as exact values whenever suitable.</i>                            |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's $d$ , Pearson's $r$ ), indicating how they were calculated  |

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection BD FACS DIVA software V8.0.1.

Data analysis FlowJo V9.9.3, V10.6.0 software(TreeStar), Wave V2.6.1 (Agilent), GraphPadPrism V7 (GraphPad Software), RStudio V1.2.1335

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

16S rDNA datasets analyzed in the manuscript are available through NCBI under accession number PRJNA528994.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- ☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. Sample sizes were determined by magnitude and consistency of measurable differences. The precise number of animals used were indicated in the figure legends.
Data exclusions	No data were excluded from analyses.
Replication	Experiments were repeated, so our data represent at least two to three independent experiments with similar results.
Randomization	Mice used in the in vivo testing of bile acids were randomly assigned to experimental groups.
Blinding	Investigators were not blinded during group allocation and data analysis. Investigators were blinded for disease scoring when performing transfer T cell colitis experimental analyses.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Antibodies

Antibodies used	Flow cytometry antibodies are purchased either from eBioscience: anti-IFN $\gamma$ (XMG1.2; #48-7311-82; Lot:1991937; 1:200); anti-IL-17a (eBio17B7; #25-7177-82; Lot:1994058; 1:200); anti-FoxP3 (FJK-16s; #11-5773-82; Lot:2007700; 1:100); anti-CD4 (RM4-5; #48-0042-82; Lot:1967921; 1:200); anti-CD3e (145-2C11; #48-0031-82; Lot:4311331; 1:200); anti-CD25 (PC61.5; #25-0251-82; Lot: E07536-1635; 1:200); anti-CD69 (H1.2F3; #45-0691-82; Lot:E08349-1633; 1:200); anti-CD62L (MEL-14; #11-0621-85; Lot:E00377-1631; 1:200); anti-Nur77(12.14; #53-5965-82; Lot:4347883; 1:100), or from Biolegend: anti-IL-4 (11B11; #504104; Lot:B271497; 1:200); anti-CD45 (30-F11; #103114; Lot:B247440; 1:200); CD45RB (C363-16A; #103308; Lot:4108938; 1:200); anti-CD44 (IM7; #103032; Lot: B238172; 1:200). Specific fluorochrome color are labeled in the figures. ChIP antibodies: anti-Rabbit IgG (#ab46540; 1:100); anti-H3K27Ac (#ab4729; Lot:GR286678-2; 1:100); anti-H3K4me1 (#ab8895; Lot:GR283603-1; 1:100); anti-P300 (#ab14984; Lot:GR272730-1; 1:100).
Validation	For eBioscience antibodies and Biolegend FACS antibodies: anti-IFN $\gamma$ (XMG1.2; #48-7311-82; Lot:1991937; 1:200); anti-IL-17a (eBio17B7; #25-7177-82; Lot:1994058; 1:200); anti-FoxP3 (FJK-16s; #11-5773-82; Lot:2007700; 1:100); anti-CD4 (RM4-5; #48-0042-82; Lot:1967921; 1:200); anti-CD3e (145-2C11; #48-0031-82; Lot:4311331; 1:200); anti-CD25 (PC61.5; #25-0251-82; Lot: E07536-1635; 1:200); anti-CD69 (H1.2F3; #45-0691-82; Lot:E08349-1633; 1:200); anti-CD62L (MEL-14; #11-0621-85; Lot:E00377-1631; 1:200); anti-Nur77(12.14; #53-5965-82; Lot:4347883; 1:100); anti-IL-4 (11B11; #504104; Lot:B271497; 1:200); anti-CD45 (30-F11; #103114; Lot:B247440; 1:200); CD45RB (C363-16A; #103308; Lot:4108938; 1:200); anti-CD44 (IM7; #103032; Lot: B238172; 1:200), manufacturer provided technical data sheets and validated antibodies for flow cytometry individually.  For ChIP antibodies: anti-Rabbit IgG (#ab46540; 1:100); anti-H3K27Ac (#ab4729; Lot:GR286678-2; 1:100); anti-H3K4me1 (#ab8895; Lot:GR283603-1; 1:100); anti-P300 (#ab14984; Lot:GR272730-1; 1:100) abcam validated each batch of antibody for ChIP analysis on its website.

## Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	293 cell line was obtained from ATCC
Authentication	293 cell line was not authenticated

Mycoplasma contamination

Cell line was not tested for mycoplasma

Commonly misidentified lines  
(See [ICLAC](#) register)

cell line was not listed in the ICLAC

## Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals

C57BL/6 (stock no. 000664), FoxP3-GFP (stock no. 016958), FXR-KO (stock no. 004144), VDR-KO (stock no. 006133), CD45.1 (stock no. 002014), PhaMexicised (stock no. 018397), Rag1-KO (stock no. 002216) mice were purchased from Jackson Laboratory. SFB containing C57BL/6 mice were purchased from Taconic bioscience. FoxP3-CNS-KO and control mice were provided by Ye Zheng Lab at Salk Institute. Both male and female mice were used in the study age range from 6-10 week old.

Wild animals

This study did not involve the use of wild animals.

Field-collected samples

This study did not involve the use of the field-collected samples.

Ethics oversight

All mouse studies were performed in full compliance with IACUC approved protocol and guidelines of Harvard Medical School.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics

This study uses fecal samples collected from patients with biopsy-proven active ulcerative colitis.

Recruitment

Eligible patients were identified by the treating physician at Weill Cornell Medicine.

Ethics oversight

Fecal samples were obtained from patients with active ulcerative colitis under an Institutional Review Board-approved protocol and informed consent was obtained at Weill Cornell Medicine IRB 1404014982.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Flow Cytometry

### Plots

Confirm that:

- ☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☒ All plots are contour plots with outliers or pseudocolor plots.
- ☒ A numerical value for number of cells or percentage (with statistics) is provided.

### Methodology

Sample preparation

This information was included in the Methods section.

Instrument

LSRII analyzer (BD Biosciences), Aria II (BD Biosciences)

Software

FACS DIVA software V8.0.1.(BD Biosciences) and FlowJo V9.9.3 software (TreeStar)

Cell population abundance

Sort-purification was carried out using FACSAria cell sorter (BD Biosciences), with >98% purity.

Gating strategy

Examples for the gating strategy were presented in Extended Data Fig.1

- ☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

# Search-and-replace genome editing without double-strand breaks or donor DNA

<https://doi.org/10.1038/s41586-019-1711-4>

Received: 26 August 2019

Accepted: 10 October 2019

Published online: 21 October 2019

Andrew V. Anzalone<sup>1,2,3</sup>, Peyton B. Randolph<sup>1,2,3</sup>, Jessie R. Davis<sup>1,2,3</sup>, Alexander A. Sousa<sup>1,2,3</sup>, Luke W. Koblan<sup>1,2,3</sup>, Jonathan M. Levy<sup>1,2,3</sup>, Peter J. Chen<sup>1,2,3</sup>, Christopher Wilson<sup>1,2,3</sup>, Gregory A. Newby<sup>1,2,3</sup>, Aditya Raguram<sup>1,2,3</sup> & David R. Liu<sup>1,2,3\*</sup>

Most genetic variants that contribute to disease<sup>1</sup> are challenging to correct efficiently and without excess byproducts<sup>2–5</sup>. Here we describe prime editing, a versatile and precise genome editing method that directly writes new genetic information into a specified DNA site using a catalytically impaired Cas9 endonuclease fused to an engineered reverse transcriptase, programmed with a prime editing guide RNA (pegRNA) that both specifies the target site and encodes the desired edit. We performed more than 175 edits in human cells, including targeted insertions, deletions, and all 12 types of point mutation, without requiring double-strand breaks or donor DNA templates. We used prime editing in human cells to correct, efficiently and with few byproducts, the primary genetic causes of sickle cell disease (requiring a transversion in *HBB*) and Tay–Sachs disease (requiring a deletion in *HEXA*); to install a protective transversion in *PRNP*; and to insert various tags and epitopes precisely into target loci. Four human cell lines and primary post-mitotic mouse cortical neurons support prime editing with varying efficiencies. Prime editing shows higher or similar efficiency and fewer byproducts than homology-directed repair, has complementary strengths and weaknesses compared to base editing, and induces much lower off-target editing than Cas9 nuclease at known Cas9 off-target sites. Prime editing substantially expands the scope and capabilities of genome editing, and in principle could correct up to 89% of known genetic variants associated with human diseases.

The ability to make virtually any targeted change in the genome of any living cell or organism is a longstanding aspiration of the life sciences. Despite rapid advances in genome editing technologies, the majority of the more than 75,000 known disease-associated genetic variants in humans<sup>1</sup> remain difficult to correct or install in most therapeutically relevant cell types (Fig. 1a). Programmable nucleases such as CRISPR–Cas9 make double-strand DNA breaks (DSBs) that can disrupt genes by inducing mixtures of insertions and deletions (indels) at target sites<sup>2–4</sup>. DSBs, however, are associated with undesired outcomes, including complex mixtures of products, translocations<sup>5</sup>, and activation of p53<sup>6,7</sup>. Moreover, the vast majority of pathogenic alleles arise from specific insertions, deletions, or base substitutions that require more precise editing technologies to correct (Fig. 1a, Supplementary Discussion). Homology-directed repair (HDR) stimulated by DSBs<sup>8</sup> has been widely used to install precise DNA changes. HDR, however, relies on exogenous donor DNA repair templates, typically generates an excess of indels from end-joining repair of DSBs, and is inefficient in most therapeutically relevant cell types (T cells and some types of stem cell being important exceptions)<sup>9,10</sup>. Whereas enhancing the efficiency and precision of DSB-mediated editing remains the focus of promising efforts<sup>11–15</sup>, these challenges motivate the exploration of alternative precision genome editing strategies.

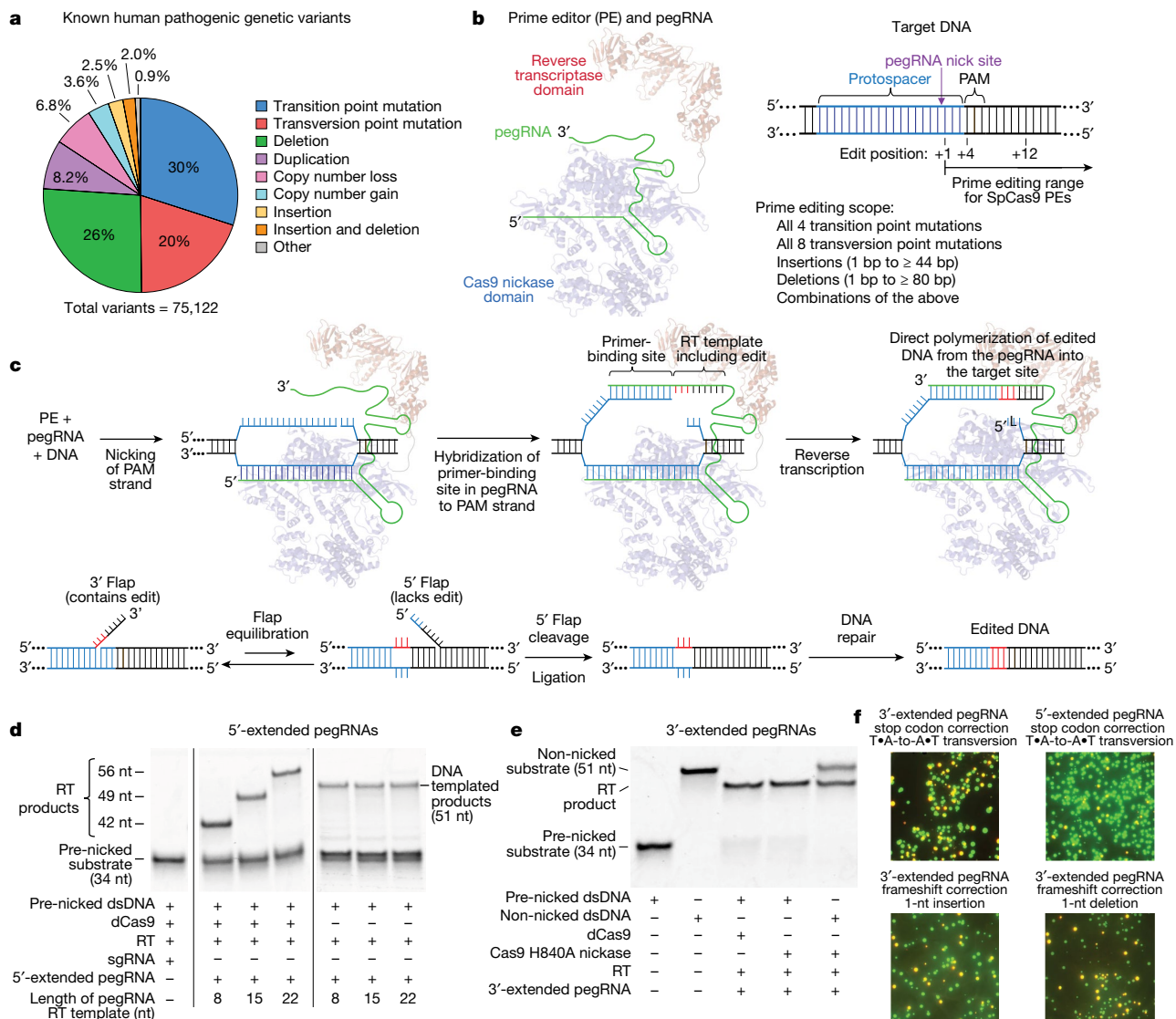
Base editing can efficiently install the four transition mutations (C→T, G→A, A→G, and T→C) without requiring DSBs in many cell types and

organisms, including mammals<sup>16–19</sup>, but cannot currently perform the eight transversion mutations (C→A, C→G, G→C, G→T, A→C, A→T, T→A, and T→G), such as the T•A-to-A•T mutation needed to directly correct the most common cause of sickle cell disease (*HBB*(E6V)). In addition, no DSB-free method has been reported to perform targeted deletions, such as the removal of the four-base duplication that causes Tay–Sachs disease (*HEXA*<sup>1278+TATC</sup>), or targeted insertions, such as the three-base insertion required to directly correct the most common cause of cystic fibrosis (*CFTR*(ΔF508)). Targeted transversions, insertions, and deletions are therefore difficult to install or correct efficiently and without excess byproducts in most cell types, even though they collectively account for most known pathogenic alleles (Fig. 1a).

Here we describe the development of prime editing, a ‘search-and-replace’ genome editing technology that mediates targeted insertions, deletions, all 12 possible base-to-base conversions, and combinations thereof in human cells without requiring DSBs or donor DNA templates. Prime editors (PEs), initially exemplified by PE1, use a reverse transcriptase (RT) fused to an RNA-programmable nickase and a prime editing guide RNA (pegRNA) to copy genetic information directly from an extension on the pegRNA into the target genomic locus. PE2 uses an engineered RT to increase editing efficiencies, while PE3 nicks the non-edited strand to induce its replacement and further increase editing efficiency, typically to 20–50% with 1–10% indel formation in human HEK293T cells. Prime

<sup>1</sup>Merkin Institute of Transformative Technologies in Healthcare, Broad Institute of Harvard and MIT, Cambridge, MA, USA. <sup>2</sup>Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA. <sup>3</sup>Howard Hughes Medical Institute, Harvard University, Cambridge, MA, USA. \*e-mail: [drliu@fas.harvard.edu](mailto:drliu@fas.harvard.edu)





**Fig. 1 | Overview of prime editing and feasibility studies in vitro and in yeast cells.** **a**, The 75,122 known pathogenic human genetic variants in ClinVar (accessed July, 2019), classified by type. **b**, A prime editing complex consists of a PE protein containing an RNA-guided DNA-nicking domain, such as Cas9 nickase, fused to an RT domain and complexed with a pegRNA. The PE–pegRNA complex enables a variety of precise DNA edits at a wide range of positions. **c**, *Streptococcus pyogenes* Cas9. **c**, The PE–pegRNA complex binds the target DNA and nicks the PAM-containing strand. The resulting 3' end hybridizes to the PBS, then primes reverse transcription of new DNA containing the desired edit using the RT template of the pegRNA. Equilibration between the edited 3' flap and the unedited 5' flap, cellular 5' flap cleavage and ligation, and DNA repair results in stably edited DNA. **d**, In vitro primer extension assays with 5'-extended pegRNAs, pre-nicked dsDNA substrates containing 5'-Cy5-

labelled PAM strands, dCas9, and a commercial M-MLV RT variant (RT, Superscript III). dCas9 was complexed with pegRNAs, then added to DNA substrates along with the indicated components. After 1 h, reactions were analysed by denaturing PAGE to visualize Cy5 fluorescence. **e**, Primer extension assays performed as in **d** using 3'-extended pegRNAs pre-complexed with dCas9 or Cas9(H840A) nickase, and pre-nicked or non-nicked dsDNA substrates. **f**, Yeast colonies transformed with GFP–mCherry fusion reporter plasmids edited in vitro with pegRNAs, Cas9 nickase, and RT. Plasmids containing nonsense or frameshift mutations between GFP and mCherry were edited with pegRNAs that restored mCherry translation via transversion, 1-bp insertion, or 1-bp deletion. GFP and mCherry double-positive cells (yellow) reflect successful editing. Images in **d–f** are representative of  $n = 2$  independent replicates. For gel source data, see Supplementary Fig. 1.

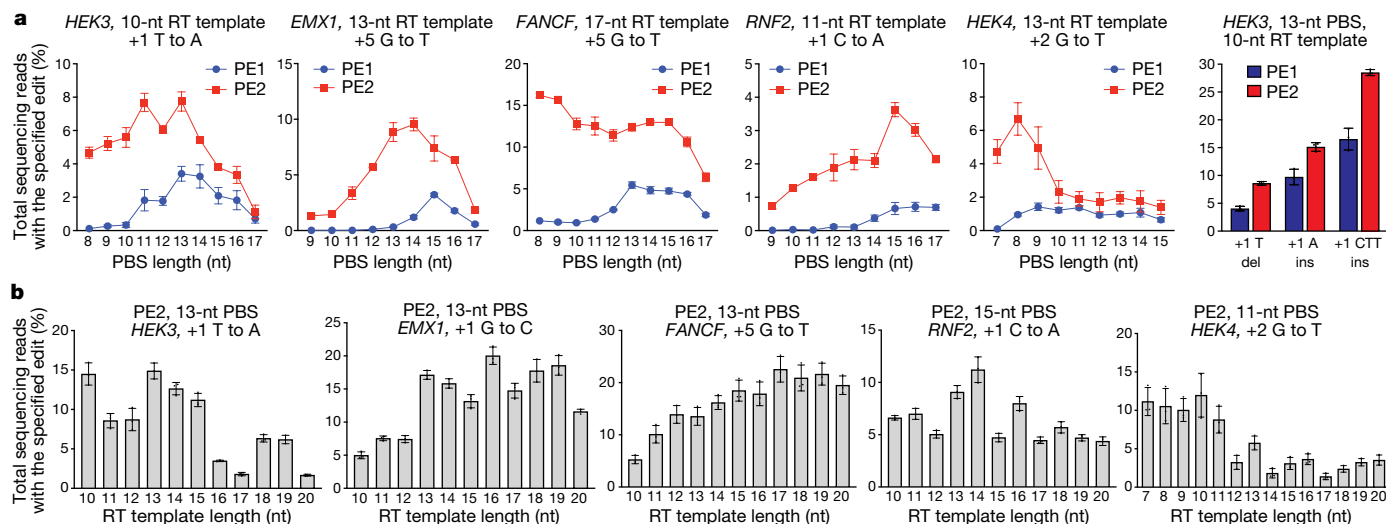
editing offers much lower off-target activity than Cas9 at known Cas9 off-target loci, far fewer byproducts and higher or similar efficiency compared to Cas9-initiated HDR, and complementary strengths and weaknesses compared to base editors. By enabling precise targeted insertions, deletions, and all 12 possible classes of point mutations without requiring DSBs or donor DNA templates, prime editing has the potential to advance the study and correction of the vast majority of pathogenic alleles.

## Prime editing strategy

Cas9 targets DNA using a guide RNA containing a spacer sequence that hybridizes to the target DNA site<sup>2–4,20,21</sup>. We envisioned the generation of

guide RNAs that both specify the DNA target and contain new genetic information that replaces target DNA nucleotides. To transfer information from these engineered guide RNAs to target DNA, we proposed that genomic DNA, nicked at the target site to expose a 3'-hydroxyl group, could be used to prime the reverse transcription of an edit-encoding extension on the engineered guide RNA (the pegRNA) directly into the target site (Fig. 1b, c, Supplementary Discussion).

These initial steps result in a branched intermediate with two redundant single-stranded DNA flaps: a 5' flap that contains the unedited DNA sequence and a 3' flap that contains the edited sequence copied from the pegRNA (Fig. 1c). Although hybridization of the perfectly complementary 5' flap to the unedited strand is likely to be thermodynamically



**Fig. 2 | Prime editing of genomic DNA in human cells by PE1 and PE2.** **a**, Use of an engineered M-MLV reverse transcriptase (D200N, L603W, T306K, W313F, T330P) in PE2 substantially improves prime editing efficiencies at five genomic sites in HEK293T cells, and small insertion and small deletion edits at *HEK3*.

**b**, PE2 editing efficiencies with varying RT template lengths at five genomic sites in HEK293T cells. Editing efficiencies reflect sequencing reads that contain the intended edit and do not contain indels among all treated cells, with no sorting. Mean  $\pm$  s.d. of  $n = 3$  independent biological replicates.

favoured, 5' flaps are the preferred substrate for structure-specific endonucleases such as FEN1<sup>22</sup>, which excises 5' flaps generated during lagging-strand DNA synthesis and long-patch base excision repair. The redundant unedited DNA may also be removed by 5' exonucleases such as EXO1<sup>23</sup>.

We reasoned that preferential 5' flap excision and 3' flap ligation could drive the incorporation of the edited DNA strand, creating heteroduplex DNA containing one edited strand and one unedited strand (Fig. 1c). DNA repair to resolve the heteroduplex by copying the information in the edited strand to the complementary strand would permanently install the edit (Fig. 1c). On the basis of a similar strategy we developed to favourably resolve heteroduplex DNA during base editing<sup>16–18</sup>, we hypothesized that nicking the non-edited DNA strand might bias DNA repair to preferentially replace the non-edited strand.

## Validation in vitro and in yeast

First, we tested whether the 3' end of the protospacer-adjacent motif (PAM)-containing DNA strand cleaved by the RuvC nuclease domain of Cas9 was sufficiently accessible to prime reverse transcription. We designed pegRNAs by adding to single guide RNAs (sgRNAs) a primer binding site (PBS) that allows the 3' end of the nicked DNA strand to hybridize to the pegRNA, and an RT template containing the desired edit (Fig. 1c). We constructed candidate pegRNAs by extending sgRNAs on either end with a PBS sequence (5–6 nucleotides (nt)) and an RT template (7–22 nt), and confirmed that 5'-extended pegRNAs support Cas9 binding to target DNA in vitro and that both 5'-extended and 3'-extended pegRNAs support Cas9-mediated DNA nicking in vitro and DNA cleavage in mammalian cells (Extended Data Fig. 1a–c). Next, we tested the compatibility of these candidate pegRNAs with reverse transcription using pre-nicked 5'-Cy5-labelled double-stranded DNA (dsDNA) substrates, catalytically dead Cas9 (dCas9), and a commercial Moloney murine leukaemia virus (M-MLV) RT variant (Extended Data Fig. 1d). When all components were present, the labelled DNA strand was efficiently converted into longer DNA products with gel mobilities consistent with reverse transcription along the RT template (Fig. 1d, Extended Data Fig. 1d, e). Omission of dCas9 led to nick translation products that resulted from RT-mediated DNA polymerization on the DNA template, with no pegRNA information transfer. No DNA polymerization products were observed when the pegRNA was replaced by

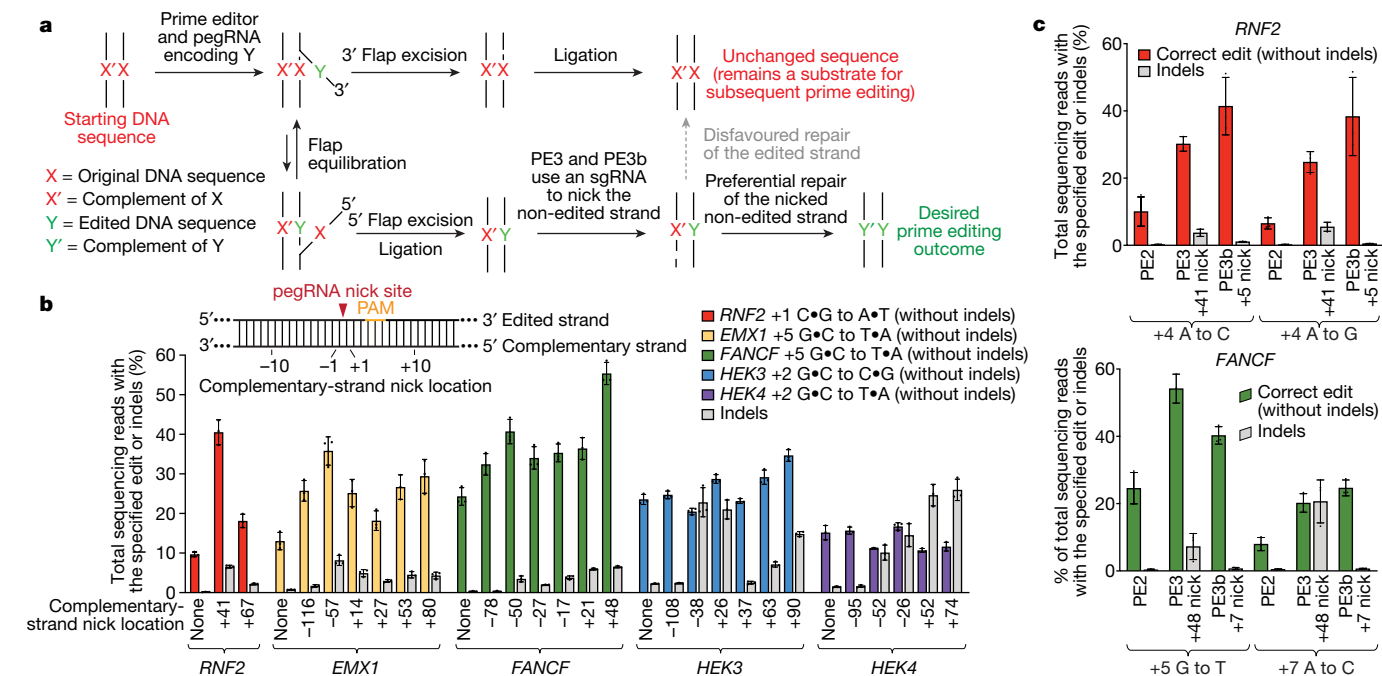
a conventional sgRNA (Fig. 1d). These results demonstrate that nicked DNA exposed by dCas9 is competent to prime reverse transcription from a pegRNA.

Next, we tested non-nicked dsDNA substrates with a Cas9(H840A) nickase that nicks the PAM-containing strand<sup>2</sup>. In these reactions, 5'-extended pegRNAs generated reverse transcription products inefficiently (Extended Data Fig. 1f), but 3'-extended pegRNAs enabled efficient Cas9 nicking and reverse transcription (Fig. 1e). The use of 3'-extended pegRNAs generated only a single apparent product, despite the theoretical possibility that reverse transcription could terminate anywhere within the pegRNA. DNA sequencing of reactions with Cas9 nickase, RT, and 3'-extended pegRNAs revealed that the complete RT template sequence was reverse transcribed into the DNA substrate (Extended Data Fig. 1g). These experiments establish that 3'-extended pegRNAs can direct Cas9 nickase and template reverse transcription in vitro.

To evaluate the eukaryotic cell DNA repair outcomes of 3' flaps produced by pegRNA-programmed reverse transcription in vitro, we performed in vitro prime editing on reporter plasmids, then transformed the reaction products into yeast cells (Extended Data Fig. 2). We constructed reporter plasmids encoding EGFP and mCherry separated by a linker containing an in-frame stop codon, +1 frameshift, or –1 frameshift. When plasmids were edited in vitro with Cas9 nickase, RT, and 3'-extended pegRNAs encoding a transversion that corrects the premature stop codon, 37% of yeast transformants expressed both GFP and mCherry (Fig. 1f, Extended Data Fig. 2). Reactions edited with 5'-extended pegRNAs yielded fewer GFP and mCherry double-positive colonies (9%). Productive editing was also observed using 3'-extended pegRNAs that insert a single nucleotide (15%) or delete a single nucleotide (29%) to correct frameshift mutations (Fig. 1f, Extended Data Fig. 2). These results demonstrate that DNA repair in eukaryotic cells can resolve 3' DNA flaps from prime editing to incorporate precise transversions, insertions, and deletions.

## Prime editor 1

Encouraged by these observations, we sought to develop a prime editing system with a minimum number of components that could edit genomic DNA in mammalian cells. We transfected HEK293T cells with one plasmid encoding a fusion of the wild-type M-MLV RT through a



**Fig. 3 | PE3 and PE3b systems nick the non-edited strand to increase prime editing efficiency.** **a**, Overview of prime editing by PE3. After initial synthesis of the edited strand, 5' flap excision leaves behind a DNA heteroduplex containing one edited strand and one non-edited strand. Mismatch repair resolves the heteroduplex to give either edited or non-edited products. Nicking the non-edited strand favours repair of that strand, resulting in preferential generation of duplex DNA containing the desired edit. **b**, The

effect of complementary strand nicking on prime editing efficiency and indel formation. 'None' refers to PE2 controls, which do not nick the complementary strand. **c**, Comparison of editing efficiencies with PE2, PE3, and PE3b (edit-specific complementary strand nick). Editing efficiencies reflect sequencing reads that contain the intended edit and do not contain indels among all treated cells, with no sorting. Mean  $\pm$  s.d. of  $n = 3$  independent biological replicates.

flexible linker to either terminus of the Cas9(H840A) nickase, and a second plasmid encoding a pegRNA (Extended Data Fig. 3a). Initial attempts led to no detectable editing.

Extension of the PBS in the pegRNA to 8–15 bases, however, led to detectable installation of a transversion at the HEK293 site 3 (hereafter referred to as HEK3) target site, with higher efficiencies when the RT was fused to the C terminus of Cas9 nickase than when it was fused to the N terminus (Extended Data Fig. 3b). These results suggest that wild-type M-MLV RT fused to Cas9 requires longer PBS sequences for genome editing in human cells compared to what is required in vitro using the commercial variant of M-MLV RT supplied in trans. We designated this M-MLV RT fused to the C terminus of Cas9(H840A) nickase as PE1.

We tested the ability of PE1 to introduce transversion point mutations at four additional genomic sites specified by the pegRNA (Fig. 2a). Editing efficiency at these sites was dependent on PBS length, with maximal editing efficiencies reaching 0.7–5.5% (Fig. 2a). Indels from PE1 were minimal, averaging  $0.2 \pm 0.1\%$  (mean  $\pm$  s.d.) for the five sites under conditions that maximized each site's editing efficiency (Extended Data Fig. 3a–f). PE1 also mediated targeted insertions and deletions with 4–17% efficiency at the HEK3 locus (Fig. 2a). These findings show that PE1 can directly install targeted transversions, insertions, and deletions without requiring DSBs or DNA templates.

## Prime editor 2

We hypothesized that engineering the RT in PE1 might improve the efficiency of DNA synthesis during prime editing. M-MLV RT mutations that increase thermostability<sup>24,25</sup>, processivity<sup>24</sup>, and DNA–RNA substrate affinity<sup>26</sup>, and that inactivate RNaseH activity<sup>27</sup>, have been reported. We constructed 19 variants of PE1 containing a variety of RT mutations to evaluate their editing efficiency in human cells.

First, we investigated M-MLV RT variants that support reverse transcription at elevated temperatures<sup>24</sup>. Introduction of D200N, L603W

and T330P into M-MLV RT, hereafter referred to as M3, led to a 6.8-fold average increase in transversion and insertion editing efficiency across five genomic loci in HEK293T cells compared to PE1 (Extended Data Fig. 4).

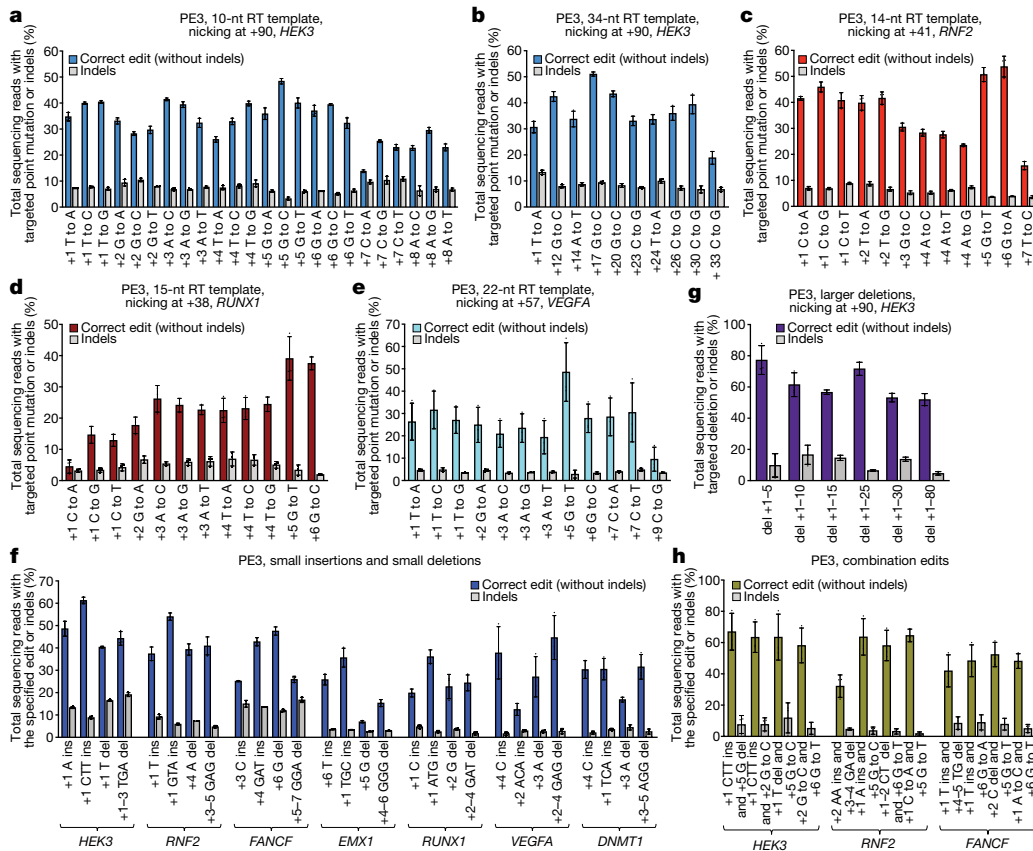
We tested additional RT mutations that have been shown to enhance binding to the template–PBS complex, enzyme processivity, and thermostability<sup>26</sup>. Among the 14 additional mutants analysed, addition of T306K and W313F to M3 improved editing efficiency an additional 1.3-fold to 3.0-fold for six transversion or insertion edits across five genomic sites (Extended Data Fig. 4). This pentamutant RT incorporated into PE1 (Cas9(H840A)–M-MLV RT(D200N/L603W/T330P/T306K/W313F)) is hereafter referred to as prime editor 2 (PE2).

PE2 installs single-nucleotide transversion, insertion, and deletion mutations with substantially higher efficiency than PE1, and is compatible with shorter PBS sequences, consistent with enhanced engagement of transient genomic DNA–PBS complexes (Fig. 2a). On average, PE2 led to a 1.6- to 5.1-fold improvement in the efficiency of prime editing point mutations over PE1. PE2 also performed targeted insertions and deletions more efficiently than PE1 (Fig. 2a, Extended Data Fig. 4d).

## Optimization of pegRNAs

We systematically probed the relationship between pegRNA structure and PE2 editing efficiency. Priming regions with lower G/C content generally required longer PBS sequences, consistent with the energetic requirements of hybridization of the nicked DNA strand to the pegRNA PBS (Fig. 2a). No PBS length or G/C content level was strictly predictive of editing efficiency, suggesting that other factors such as DNA primer or RT template secondary structure also influence editing activity. We recommend starting with a PBS length of about 13 nt, and testing different PBS lengths during optimization, especially if the priming region deviates from about 40–60% G/C content.





**Fig. 4 | Targeted insertions, deletions, and all 12 types of point mutation with PE3 at seven endogenous genomic loci in HEK293T cells. a**, All 12 types of single-nucleotide edit from position +1 to +8 of the *HEK3* site using a 10-nt RT template, counting the first nucleotide following the pegRNA-induced nick as position +1. **b**, Long-range PE3 edits at *HEK3* using a 34-nt RT template. **c–e**, PE3-mediated transition and transversion edits at the specified positions for *RNF2*

(c), *RUNX1* (d), and *VEGFA* (e). **f**, Targeted 1- and 3-bp insertions, and 1- and 3-bp deletions with PE3 at seven endogenous genomic loci. **g**, Targeted precise deletions of 5–80 bp at *HEK3*. **h**, Combination edits at three endogenous genomic loci. Editing efficiencies reflect sequencing reads that contain the intended edit and do not contain indels among all treated cells, with no sorting. Mean  $\pm$  s.d. of  $n = 3$  independent biological replicates.

Next, we systematically evaluated pegRNAs with RT templates 10–20 nt long at five genomic target sites using PE2 (Fig. 2b), and with RT templates up to 31 nt at three genomic sites (Extended Data Fig. 5a–c). As with PBS length, RT template length could also be varied to maximize prime editing efficiency, although many RT template lengths of ten or more nucleotides performed comparably. As some target sites preferred longer RT templates (more than 15 nt; *FANCF*, *EMX1*), whereas other loci preferred shorter RT templates (*HEK3* and *HEK293* site 4, hereafter referred to as *HEK4*) (Fig. 2b), we recommend starting with about 10–16 nt and testing shorter and longer RT templates during pegRNA optimization.

Notably, the use of RT templates that place a C adjacent to the 3' hairpin of the sgRNA scaffold generally resulted in lower editing efficiency (Extended Data Fig. 5a–c). We speculate that a C as the first nucleotide of the 3' extension can disrupt guide RNA structure by pairing with G81, which normally forms a pi stack with Y1356 in Cas9 and a non-canonical base pair with A68 of the sgRNA<sup>28</sup>. Because many RT template lengths support prime editing, we recommend designing pegRNAs so that the first base of the 3' extension is not C.

## Prime editor 3 systems

The resolution of heteroduplex DNA from PE2 containing one edited and one non-edited strand determines long-term editing outcomes. To optimize base editing we previously used Cas9 nickase to nick the non-edited strand, directing DNA repair to that strand using the edited strand as a template<sup>16–18</sup>. To apply this strategy to enhance prime editing, we tested nicking the non-edited strand using the Cas9(H840A)

nickase already present in PE2 and a simple sgRNA (Fig. 3a). As the edited DNA strand is also nicked to initiate prime editing, we tested a variety of nick locations on the non-edited strand to minimize DSBs that lead to indels.

We first tested this strategy, designated PE3, at five genomic sites in HEK293T cells using sgRNAs that induce nicks 14–116 nt away from the site of the pegRNA-induced nick. In four of the five sites tested, nicking the non-edited strand increased editing efficiency by 1.5- to 4.2-fold compared to PE2, to as high as 55% (Fig. 3b). Although the optimal nicking position varied depending on the genomic site (Supplementary Discussion), nicks positioned 3' of the edit about 40–90 bp from the pegRNA-induced nick generally increased editing efficiency (averaging 41%) without excess indel formation (6.8% average indels for the sgRNA with the highest editing efficiency) (Fig. 3b). We recommend starting with non-edited strand nicks about 50 bp from the pegRNA-mediated nick, and testing alternative nick locations if indel frequencies exceed acceptable levels.

Nicking the non-edited strand only after resolution of the edited strand flap should minimize the presence of concurrent nicks, thereby minimizing formation of DSBs and indels. To achieve this goal, we designed sgRNAs with spacers that matched the edited strand, but not the original allele. Using this strategy, denoted PE3b, mismatches between the spacer and the unedited allele should disfavor sgRNA nicking until after editing of the PAM strand has taken place. PE3b resulted in a 13-fold decrease in the average number of indels (0.74%) compared to PE3, without any evident decrease in editing efficiency (Fig. 3c). When the edit lies within a second protospacer, we recommend the PE3b approach.

Together, these findings establish that PE3 systems improve editing efficiencies about threefold compared with PE2, albeit with a higher range of indels than PE2. When it is possible to nick the non-edited strand with an sgRNA that requires editing before nicking, the PE3b system offers PE3-like editing levels while greatly reducing indel formation.

To demonstrate the targeting scope and versatility of prime editing with PE3, we performed all 24 possible single-nucleotide substitutions across the +1 to +8 positions (counting the first base 3' of the pegRNA-induced nick as position +1) of the *HEK3* target site using PE3 and pegRNAs with 10-nt RT templates (Fig. 4a). These 24 edits collectively cover all 12 possible transition and transversion mutations, and proceeded with average editing efficiencies (containing no indels) of  $33 \pm 7.9\%$ , with  $7.5 \pm 1.8\%$  average indels.

Notably, long-distance RT templates can also give rise to efficient prime editing. Using PE3 with a 34-nt RT template, we installed point mutations at positions +12, +14, +17, +20, +23, +24, +26, +30, and +33 in the *HEK3* locus with  $36 \pm 8.7\%$  average efficiency and  $8.6 \pm 2.0\%$  indels (Fig. 4b). Other RT templates of 30 or more nucleotides at three other genomic sites also supported prime editing (Extended Data Fig. 5a–c). As an NGG PAM on either DNA strand occurs on average every 8 bp, far less than edit-to-PAM distances that support efficient prime editing, prime editing is not substantially constrained by the availability of a nearby PAM sequence, in contrast to other precision editing methods<sup>11,15,16</sup>. Given the presumed relationship between RNA secondary structure and prime editing efficiency, when designing pegRNAs for long-range edits we recommend testing RT templates of various lengths and, if necessary, sequence compositions (for example, using synonymous codons).

To further test the scope and limitations of PE3 for introducing point mutations, we tested 72 additional edits covering all possible types of point mutation across six additional genomic target sites (Fig. 4c–e, Extended Data Fig. 5d–f). Editing efficiency averaged  $25 \pm 14\%$ , while indel formation averaged  $8.3 \pm 7.5\%$ . Because the pegRNA RT template includes the PAM sequence, prime editing can induce changes in the PAM sequence. In these cases, we observed higher editing efficiency (averaging  $39 \pm 9.7\%$ ) and lower indel generation (averaging  $5.0 \pm 2.9\%$ ; Fig. 4, mutations at +5 or +6), potentially due to the inability of Cas9 nickase to re-bind and nick the edited strand before the repair of the complementary strand. We recommend editing the PAM, in addition to other desired changes, whenever possible.

Next, we performed 28 targeted small insertions and small deletions at seven genomic sites using PE3 (Fig. 4f). Targeted 1-bp and 3-bp insertions proceeded with an average efficiency of  $32 \pm 9.8\%$  and  $39 \pm 16\%$ , respectively. Targeted 1-bp and 3-bp deletions were also efficient, averaging  $29 \pm 14\%$  and  $32 \pm 11\%$  editing, respectively. Indel generation (beyond the target insertion or deletion) averaged  $6.8 \pm 5.4\%$ . Because insertions and deletions between positions +1 and +6 alter the location or structure of the PAM, we speculate that insertions or deletions at these positions are more efficient because they prevent re-engagement of the edited strand.

We also tested PE3 for its ability to mediate larger precise deletions of 5–80 bp at the *HEK3* site (Fig. 4g). We observed very high editing efficiencies (52–78%) for precise 5-, 10-, 15-, 25-, and 80-bp deletions, with indels averaging  $11 \pm 4.8\%$ . Finally, we tested the ability of PE3 to mediate 12 combinations of insertions, deletions, and/or point mutations across three genomic sites. These combination edits were also very efficient, averaging 55% editing with 6.4% indels (Fig. 4h). Together, the 156 distinct edits in Fig. 4 and Extended Data Fig. 5d–f establish the versatility, precision, and targeting flexibility of PE3 systems.

### Prime editing compared with base editing

Cytidine base editors (CBEs) and adenine base editors (ABEs) can install transition mutations efficiently and with few indels<sup>16–18</sup>. The application of base editing can be limited by unwanted bystander edits from

the presence of multiple cytidine or adenine bases within the base editing activity window<sup>16–18,29</sup>, or by the absence of a PAM positioned about  $15 \pm 2$  nt from the target nucleotide<sup>16,30</sup>. We anticipated that prime editing could complement base editing when bystander edits are unacceptable or when the target site lacks a suitably positioned PAM.

We compared PEs and CBEs at three genomic loci that contain multiple target cytosines in the canonical base editing window (protospacer positions 4–8, counting the PAM as positions 21–23) using current-generation CBEs<sup>31</sup> without or with nickase activity (BE2max and BE4max, respectively), or using analogous PE2 and PE3 prime editing systems. Among the nine total cytosines within the base editing windows of the three sites, BE4max yielded 2.2-fold higher average total C•G-to-T•A conversion than PE3 for bases in the centre of the base editing window (protospacer positions 5–7, Extended Data Fig. 6a). However, PE3 outperformed BE4max by 2.7-fold at cytosines positioned outside the centre of the base editing window. Overall, indel frequencies for PE2 were very low (averaging  $0.86 \pm 0.47\%$ ), and for PE3 were similar to or modestly higher than that of BE4max (PE3: 2.5–21%; BE4max: 2.5–14%) (Extended Data Fig. 6b).

For the installation of precise edits (with no bystander editing), the efficiency of prime editing greatly exceeded that of base editing at the above sites, which, like most genomic DNA sites, contain multiple cytosines within the base editing window. BE4max generated few products containing only the single target base-pair conversion with no bystander edits. By contrast, prime editing at this site could be used to selectively install a C•G-to-T•A edit at any position or combination of positions (Extended Data Fig. 6c).

We also compared nicking and non-nicking adenine base editors (ABEs) with PE3 and PE2, with similar results (Extended Data Fig. 6d–f, Supplementary Discussion). Collectively, these results indicate that base editing and prime editing offer complementary strengths and weaknesses for making targeted transition mutations. When a single target nucleotide is present within the base editing window, or when bystander edits are acceptable, current base editors are typically more efficient and generate fewer indels than prime editors. When multiple cytosines or adenines are present and bystander edits are undesirable, or when PAMs that position target nucleotides for base editing are not available, prime editors offer substantial advantages.

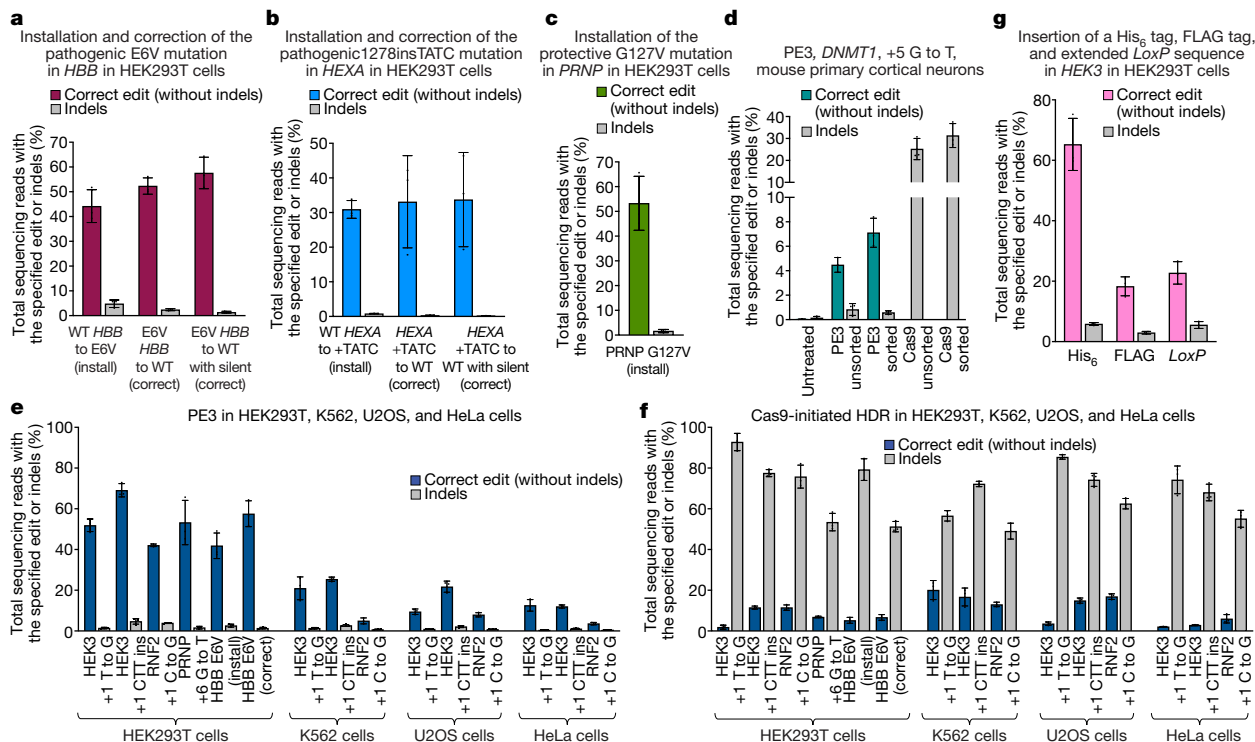
### Off-target prime editing

Prime editing requires target DNA–pegRNA spacer complementarity for the Cas9 domain to bind, target DNA–pegRNA PBS complementarity to initiate pegRNA-templated reverse transcription, and target DNA–RT product complementarity for flap resolution. To test whether these three distinct DNA hybridization steps reduce off-target prime editing compared to editing methods that require only target–guide RNA complementarity, we treated HEK293T cells with PE3 or PE2 and 16 pegRNAs that target four genomic loci, each of which has at least four well-characterized Cas9 off-target sites<sup>32,33</sup>. We also treated cells with Cas9 nuclease and the same 16 pegRNAs, or with Cas9 and four sgRNAs targeting the same four protospacers (Supplementary Table 1).

Consistent with previous studies<sup>32</sup>, Cas9 and sgRNAs targeting *HEK3*, *HEK4*, *EMX1*, and *FANCF* modified the top four known Cas9 off-target loci for each sgRNA with average frequencies of  $16 \pm 16\%$ ,  $60 \pm 26\%$ ,  $48 \pm 28\%$ , and  $4.3 \pm 5.6\%$ , respectively (Extended Data Fig. 6g). Cas9 with pegRNAs modified on-target sites with similar efficiency as Cas9 with sgRNAs, whereas Cas9 with pegRNAs modified off-target sites at 4.4-fold lower average efficiency than Cas9 with sgRNAs.

Strikingly, PE3 or PE2 with the same 16 pegRNAs containing these four target spacers resulted in detectable off-target editing at only 3 out of 16 off-target sites, with only 1 of 16 showing an off-target editing efficiency of 1% or more (Extended Data Fig. 6h). Average off-target prime editing for pegRNAs targeting *HEK3*, *HEK4*, *EMX1*, and *FANCF* at the top four known Cas9 off-target sites for each protospacer was





**Fig. 5 | Prime editing of pathogenic mutations, prime editing in primary mouse cortical neurons, and comparison of prime editing and HDR in four human cell lines. a**, Installation (via T•A-to-A•T transversion) and correction (via A•T-to-T•A transversion) of the pathogenic E6V-coding mutation in *HBB* in HEK293T cells. Correction either to wild-type *HBB*, or to *HBB* containing a PAM-disrupting silent mutation, is shown. **b**, Installation (via 4-bp insertion) and correction (via 4-bp deletion) of the pathogenic *HEXA*<sup>1278insTATC</sup> allele in HEK293T cells. Correction either to wild-type *HEXA*, or to *HEXA* containing a PAM-disrupting silent mutation, is shown. **c**, Installation of the protective G127V-coding variant in *PRNP* in HEK293T cells via G•C-to-T•A transversion. **d**,

Installation of a G•C-to-T•A transversion in *DNMT1* of mouse primary cortical neurons using a split-intein PE3 lentivirus system (see Methods). Sorted values reflect editing or indels from GFP-positive nuclei, while unsorted values are from all nuclei. **e**, **f**, PE3 editing and indels (**e**) or Cas9-initiated HDR editing and indels (**f**) at endogenous genomic loci in HEK293T, K562, U2OS, and HeLa cells. **g**, Targeted insertion of a His<sub>6</sub> tag (18 bp), Flag epitope tag (24 bp), or extended *loxP* site (44 bp) in HEK293T cells by PE3. Editing efficiencies reflect sequencing reads that contain the intended edit and do not contain indels among all treated cells, with no sorting, except where specified in **e**. Mean  $\pm$  s.d. of  $n=3$  independent biological replicates.

<0.1%, <2.2  $\pm$  5.2%, <0.1%, and <0.13  $\pm$  0.11%, respectively (Extended Data Fig. 6h). Notably, at the *HEK4* off-target 3 site that was edited by Cas9 with pegRNA1 at 97% efficiency, PE2 with pegRNA1 resulted in only 0.2% off-target editing despite sharing the same pegRNA, demonstrating how the two additional hybridization events required for prime editing can greatly reduce off-target modification. Together, these results suggest that prime editing induces much lower off-target editing than Cas9 at known Cas9 off-target sites.

Reverse transcription of 3'-extended pegRNAs in principle can proceed into the guide RNA scaffold, resulting in scaffold sequence insertion that contributes to indels at the target locus. We analysed 66 PE3 editing experiments at four loci in HEK293T cells and observed 1.7  $\pm$  1.5% average total insertion of any number of pegRNA scaffold nucleotides (Extended Data Fig. 7). We speculate that inaccessibility of the guide RNA scaffold to reverse transcription due to Cas9 domain binding, and cellular excision of the mismatched 3' end of 3' flaps that extend into the pegRNA scaffold, minimize products that incorporate pegRNA scaffold nucleotides.

The presence of endogenous human RTs from retroelements<sup>34</sup> and telomerase suggests that RT activity is not inherently toxic to human cells. Indeed, we observed no differences in the viability of HEK293T cells expressing dCas9, Cas9(H840A) nickase, PE2, or PE2 with R110S and K103L mutations (PE2-dRT) that inactivate the RT and abolish prime editing<sup>35</sup> (Extended Data Fig. 8a, b). To evaluate changes in the cellular transcriptome that result from prime editing, we performed RNA sequencing (RNA-seq) on HEK293T cells expressing PE2, PE2-dRT, or Cas9(H840A) nickase together with a *PRNP*-targeting or *HEXA*-targeting pegRNA (Extended Data Fig. 8c–k), and observed that active

PE2 minimally perturbed the transcriptome relative to Cas9 nickase or a control lacking active RT (Supplementary Discussion).

## Prime editing pathogenic mutations

We tested the ability of PE3 to directly install or correct in human cells transversion, insertion, and deletion mutations that cause genetic diseases. Sickle cell disease is caused by a A•T-to-T•A transversion mutation in *HBB*, resulting in an E6V mutation in  $\beta$ -globin (Supplementary Discussion). We used PE3 to install this *HBB* mutation into HEK293T cells with 44% efficiency and 4.8% indels (Fig. 5a) and isolated from a single prime editing experiment six HEK293T cell lines that were homozygous (triploid) for the mutated *HBB* allele (Supplementary Note 1). To correct the mutant *HBB* allele to wild-type *HBB*, we treated HEK293T cells homozygous for mutant *HBB* with PE3 and a pegRNA programmed to directly revert the *HBB* mutation to wild-type *HBB*. All 14 tested pegRNAs mediated efficient correction of mutant *HBB* to wild-type *HBB* (26–52% efficiency), and indel levels averaged 2.8  $\pm$  0.70% (Extended Data Fig. 9a). Introduction of a PAM-modifying silent mutation improved editing efficiency and product purity to 58% correction with 1.4% indels (Fig. 5a).

The most common mutation that causes Tay-Sachs disease is a 4-bp insertion in *HEXA* (*HEXA*<sup>1278insTATC</sup>). We used PE3 to install this 4-bp insertion into *HEXA* with 31% efficiency and 0.8% indels (Fig. 5b), and isolated two HEK293T cell lines that were homozygous for *HEXA*<sup>1278insTATC</sup> (Supplementary Note 1). We used these cells to test 43 pegRNAs and three nicking sgRNAs with PE3 or PE3b systems for correction of the pathogenic insertion in *HEXA* (Extended Data Fig. 9b). Nineteen of the

43 pegRNAs tested resulted in editing with an efficiency of 20% or more. Correction to wild-type *HEXA* with the best pegRNA proceeded with 33% efficiency and 0.32% indels using PE3b (Fig. 5b, Extended Data Fig. 9b).

Finally, we used PE3 to install a protective G•C-to-T•A transversion into *PRNP* (resulting in PRNP(G127V)) into HEK293T cells, introducing a mutant allele that confers resistance to prion disease in humans<sup>36</sup> and mice<sup>37</sup> (Supplementary Discussion). We evaluated four pegRNAs and three nicking sgRNAs. The most effective pegRNA with PE3 resulted in 53% installation of G127V, with 1.7% indels (Fig. 5c). Together, these results establish the ability of prime editing in human cells to install or correct transversion, insertion, or deletion mutations that cause or confer resistance to disease efficiently, and with few byproducts.

## Other cell lines and primary neurons

Next, we tested prime editing at endogenous sites in three additional human cell lines (Extended Data Fig. 10a, Supplementary Discussion). In K562 cells, PE3 achieved three transversion edits and a His<sub>6</sub> tag insertion with 15–30% editing efficiency and 0.85–2.2% indels (Extended Data Fig. 10a). In U2OS cells, we installed transversion mutations, as well as a 3-bp insertion and His<sub>6</sub> tag insertion, with 7.9–22% editing efficiency and 0.13–2.2% indels (Extended Data Fig. 10a). Finally, in HeLa cells we performed a 3-bp insertion with 12% average efficiency and 1.3% indels (Extended Data Fig. 10a). Collectively, these data indicate that cell lines other than HEK293T support prime editing, although editing efficiencies vary by cell type and are generally less efficient than in HEK293T cells. Editing:indel ratios remained favourable in all human cell lines tested.

To determine whether prime editing is possible in post-mitotic, terminally differentiated primary cells, we transduced primary cortical neurons from E18.5 mice with a PE3 lentiviral delivery system in which PE2 protein components were expressed from the neuron-specific synapsin promoter<sup>38</sup> along with a GFP marker (see Methods). Nuclei were isolated two weeks after transduction and sequenced directly, or sorted for GFP expression before sequencing. We observed 7.1% average prime editing of *DNMT1* with 0.58% average indels in sorted cortical neuron nuclei (Fig. 5d). Cas9 nuclease in the same lentivirus system resulted in 31% average indels among sorted nuclei (Fig. 5d). These data indicate that post-mitotic, terminally differentiated primary cells can support prime editing.

## Prime editing compared with HDR

Finally, we compared the performance of PE3 with that of optimized Cas9-initiated HDR<sup>11,14</sup> in mitotic cell lines that support HDR<sup>14</sup>. We treated HEK293T, HeLa, K562 and U2OS cells with Cas9 nuclease, an sgRNA, and a single-stranded DNA (ssDNA) donor template designed to install a variety of transversion and insertion edits (Fig. 5e, f, Extended Data Fig. 10). Cas9-initiated HDR in all cases successfully installed the desired edit, but with far higher levels of indel byproducts than with PE3, as expected given that Cas9 induces DSBs. In HEK293T cells, the ratio of editing to indels for installation or correction of the allele encoding HBB(E6V) or installation of the allele encoding PRNP(G127V) was on average 270-fold higher for PE3 than for Cas9-initiated HDR.

Comparisons between PE3 and HDR in human cell lines other than HEK293T showed similar results, although with lower PE3 editing efficiencies (Fig. 5e, f, Supplementary Discussion). Collectively, these data indicate that HDR typically results in similar or lower editing efficiencies than PE3 with far more indels in four tested human cell lines (Extended Data Fig. 10).

## Discussion and future directions

The ability to insert arbitrary DNA sequences with single-nucleotide precision is an especially promising capability of prime editing. For

example, we used PE3 in HEK293T cells to precisely insert into *HEK3* a His<sub>6</sub> tag (18 bp, 65% efficiency), a Flag epitope tag (24 bp, 18% efficiency), and an extended Cre recombinase *loxP* site (44 bp, 23% efficiency) with 3.0–5.9% indels (Fig. 5g). We anticipate that the ability to efficiently and precisely insert new DNA sequences into target sites in living cells will enable many biotechnological and therapeutic applications.

Collectively, the prime editing experiments described here performed 19 insertions up to 44 bp, 23 deletions up to 80 bp, 119 point mutations including 83 transversions, and 18 combination edits at 12 endogenous loci in the human and mouse genomes at locations ranging from 3 bp upstream to 29 bp downstream of a PAM without making explicit DSBs. These results establish prime editing as a remarkably versatile genome editing method. Because 85–99% of insertions, deletions, indels, and duplications in ClinVar are 30 bp in length or smaller (Extended Data Fig. 11), in principle prime editing could correct up to about 89% of the 75,122 pathogenic human genetic variants in ClinVar (Fig. 1a).

Prime editing offers many possible choices of pegRNA-induced nick locations, sgRNA-induced second nick locations, PBS lengths, RT template lengths, and which strand to edit first. This flexibility, which contrasts with more limited options typically available for other precision editing methods<sup>11,15,16</sup>, allows editing efficiency, product purity, DNA specificity, and other parameters to be optimized to suit a given application (Extended Data Fig. 9).

Much additional research is needed to further understand and improve prime editing in a broad range of cell types and organisms, to assess off-target prime editing in a genome-wide manner, and to further characterize the extent to which prime editors might affect cells. Interfacing prime editing with additional in vitro and in vivo delivery strategies is essential for exploring the potential of prime editing to enable applications, including the study and treatment of genetic diseases. By enabling precise targeted transitions, transversions, insertions, and deletions in the genomes of mammalian cells without requiring DSBs, donor DNA templates, or HDR, however, prime editing provides a new search-and-replace capability that substantially expands the scope of genome editing.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1711-4>.

- Landrum, M. J. et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).
- Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
- Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
- Mali, P. et al. RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
- Kosicki, M., Tomberg, K. & Bradley, A. Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements. *Nat. Biotechnol.* **36**, 765–771 (2018).
- Haapaniemi, E., Botla, S., Persson, J., Schmierer, B. & Taipale, J. CRISPR–Cas9 genome editing induces a p53-mediated DNA damage response. *Nat. Med.* **24**, 927–930 (2018).
- Ihry, R. J. et al. p53 inhibits CRISPR–Cas9 engineering in human pluripotent stem cells. *Nat. Med.* **24**, 939–946 (2018).
- Rouet, P., Smih, F. & Jasin, M. Expression of a site-specific endonuclease stimulates homologous recombination in mammalian cells. *Proc. Natl Acad. Sci. USA* **91**, 6064–6068 (1994).
- Chapman, J. R., Taylor, M. R. G. & Boulton, S. J. Playing the end game: DNA double-strand break repair pathway choice. *Mol. Cell* **47**, 497–510 (2012).
- Cox, D. B. T., Platt, R. J. & Zhang, F. Therapeutic genome editing: prospects and challenges. *Nat. Med.* **21**, 121–131 (2015).
- Paquet, D. et al. Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. *Nature* **533**, 125–129 (2016).
- Chu, V. T. et al. Increasing the efficiency of homology-directed repair for CRISPR–Cas9-induced precise gene editing in mammalian cells. *Nat. Biotechnol.* **33**, 543–548 (2015).

13. Maruyama, T. et al. Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining. *Nat. Biotechnol.* **33**, 538–542 (2015).
14. Rees, H. A., Yeh, W.-H. & Liu, D. R. Development of hRad51-Cas9 nickase fusions that mediate HDR without double-stranded breaks. *Nat. Commun.* **10**, 2212 (2019).
15. Shen, M. W. et al. Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature* **563**, 646–651 (2018).
16. Rees, H. A. & Liu, D. R. Base editing: precision chemistry on the genome and transcriptome of living cells. *Nat. Rev. Genet.* **19**, 770–788 (2018).
17. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).
18. Gaudelli, N. M. et al. Programmable base editing of A-T to G-C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017).
19. Gao, X. et al. Treatment of autosomal dominant hearing loss by *in vivo* delivery of genome editing agents. *Nature* **553**, 217–221 (2018).
20. Marraffini, L. A. & Sontheimer, E. J. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**, 1843–1845 (2008).
21. Barrangou, R. et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
22. Liu, Y., Kao, H.-I. & Bambara, R. A. Flap endonuclease 1: a central component of DNA metabolism. *Annu. Rev. Biochem.* **73**, 589–615 (2004).
23. Keijzers, G., Bohr, V. A. & Rasmussen, L. J. Human exonuclease 1 (EXO1) activity characterization and its function on flap structures. *Biosci. Rep.* **35**, e00206 (2015).
24. Baranauskas, A. et al. Generation and characterization of new highly thermostable and processive M-MuLV reverse transcriptase variants. *Protein Eng. Des. Sel.* **25**, 657–668 (2012).
25. Gerard, G. F. et al. The role of template-primer in protection of reverse transcriptase from thermal inactivation. *Nucleic Acids Res.* **30**, 3118–3129 (2002).
26. Arezi, B. & Hogrefe, H. Novel mutations in Moloney murine leukemia virus reverse transcriptase increase thermostability through tighter binding to template-primer. *Nucleic Acids Res.* **37**, 473–481 (2009).
27. Kotewicz, M. L., Sampson, C. M., D'Alessio, J. M. & Gerard, G. F. Isolation of cloned Moloney murine leukemia virus reverse transcriptase lacking ribonuclease H activity. *Nucleic Acids Res.* **16**, 265–277 (1988).
28. Nishimasu, H. et al. Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935–949 (2014).
29. Thuronyi, B. W. et al. Continuous evolution of base editors with expanded target compatibility and improved activity. *Nat. Biotechnol.* **37**, 1070–1079 (2019).
30. Kim, Y. B. et al. Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. *Nat. Biotechnol.* **35**, 371–376 (2017).
31. Koblan, L. W. et al. Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nat. Biotechnol.* **36**, 843–846 (2018).
32. Tsai, S. Q. et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* **33**, 187–197 (2015).
33. Kleinstiver, B. P. et al. High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).
34. Bannert, N. & Kurth, R. Retroelements and the human genome: new perspectives on an old relation. *Proc. Natl Acad. Sci. USA* **101** (Suppl. 2), 14572–14579 (2004).
35. Halvas, E. K., Svarovskaia, E. S. & Pathak, V. K. Role of murine leukemia virus reverse transcriptase deoxyribonucleoside triphosphate-binding site in retroviral replication and *in vivo* fidelity. *J. Virol.* **74**, 10349–10358 (2000).
36. Mead, S. et al. A novel protective prion protein variant that colocalizes with kuru exposure. *N. Engl. J. Med.* **361**, 2056–2065 (2009).
37. Asante, E. A. et al. A naturally occurring variant of the human prion protein completely prevents prion disease. *Nature* **522**, 478–481 (2015).
38. Kügler, S., Kilic, E. & Bähr, M. Human synapsin 1 gene promoter confers highly neuron-specific long-term transgene expression from an adenoviral vector in the adult rat brain depending on the transduced area. *Gene Ther.* **10**, 337–347 (2003).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

# Article

## Methods

### General methods

DNA amplification was conducted by PCR using Phusion U Green Multiplex PCR Master Mix (ThermoFisher Scientific) or Q5 Hot Start High-Fidelity 2× Master Mix (New England Biolabs) unless otherwise noted. DNA oligonucleotides, including Cy5-labelled DNA oligonucleotides, dCas9 protein, and Cas9(H840A) protein were obtained from Integrated DNA Technologies. Yeast reporter plasmids were derived from previously described plasmids<sup>39</sup> and cloned by the Gibson assembly method. All mammalian editor plasmids used in this work were assembled using the USER cloning method as previously described<sup>40</sup>. Plasmids expressing sgRNAs were constructed by ligation of annealed oligonucleotides into BsmBI-digested acceptor vector (Addgene plasmid no. 65777). Plasmids expressing pegRNAs were constructed by Gibson assembly or Golden Gate assembly using a custom acceptor plasmid (see Supplementary Note 3). Sequences of sgRNA and pegRNA constructs used in this work are listed in Supplementary Tables 2 and 3. All vectors for mammalian cell experiments were purified using Plasmid Plus Midiprep kits (Qiagen) or PureYield plasmid miniprep kits (Promega), which include endotoxin removal steps. All experiments using live animals were approved by the Broad Institute Institutional and Animal Care and Use Committees. Wild-type C57BL/6 mice were obtained from Charles River (#027). No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

### In vitro biochemical assays

pegRNAs and sgRNAs were transcribed in vitro using the HiScribe T7 in vitro transcription kit (New England Biolabs) from PCR-amplified templates containing a T7 promoter sequence. RNA was purified by denaturing urea PAGE and quality-confirmed by an analytical gel before use. 5'-Cy5-labelled DNA duplex substrates were annealed using two oligonucleotides (Cy5-AVA024 and AVA025; 1:1.1 ratio) for the non-nicked substrate or three oligonucleotides (Cy5-AVA023, AVA025 and AVA026; 1:1.1:1.1) for the pre-nicked substrate by heating to 95 °C for 3 min followed by slowly cooling to room temperature (Supplementary Table 2). Cas9 cleavage and reverse transcription reactions were carried out in 1× cleavage buffer<sup>41</sup> supplemented with dNTPs (20 mM HEPES-K, pH 7.5; 100 mM KCl; 5% glycerol; 0.2 mM EDTA, pH 8.0; 3 mM MgCl<sub>2</sub>; 0.5 mM dNTP mix; 5 mM DTT). dCas9 or Cas9(H840A) (5 μM final) and the sgRNA or pegRNA (5 μM final) were pre-incubated at room temperature in a 5-μl reaction mixture for 10 min before the addition of 0.5 μl of 4 μM duplex DNA substrate (400 nM final), followed by the addition of 0.2 μl of Superscript III reverse transcriptase (ThermoFisher Scientific), an undisclosed M-MLV RT variant, when applicable. Reactions were carried out at 37 °C for 1 h, then diluted to a volume of 10 μl with water, treated with 0.2 μl of proteinase K solution (20 mg/ml, ThermoFisher Scientific), and incubated at room temperature for 30 min. Following heat inactivation at 95 °C for 10 min, reaction products were combined with 2× formamide gel loading buffer (90% formamide; 10% glycerol; 0.01% bromophenol blue), denatured at 95 °C for 5 min, and separated by denaturing urea PAGE gel (15% TBE-urea, 55 °C, 200 V). DNA products were visualized by Cy5 fluorescence signal using a Typhoon FLA 7000 biomolecular imager.

Electrophoretic mobility shift assays were carried out in 1× binding buffer (1× cleavage buffer with 10 μg/ml heparin) using pre-incubated dCas9–sgRNA or dCas9–pegRNA complexes (concentration between 5 nM and 1 μM final) and Cy5-labelled duplex DNA (Cy5-AVA024 and AVA025; 20 nM final). After 15 min of incubation at 37 °C, the samples were analysed by native PAGE gel (10% TBE) and imaged for Cy5 fluorescence.

For DNA sequencing of reverse transcription products, fluorescent bands were excised and purified from urea PAGE gels, then 3' tailed with terminal transferase (TdT; New England Biolabs) in the presence of dGTP or dATP according to the manufacturer's protocol. Tailed DNA

products were diluted tenfold with binding buffer (40% saturated aqueous guanidinium chloride and 60% isopropanol) and purified by QIAquick spin column (Qiagen), then used as templates for primer extension by Klenow fragment (New England Biolabs) using primer AVA134 (A-tailed products) or AVA135 (G-tailed products) (Supplementary Table 2). Extensions were amplified by PCR for 10 cycles using primers AVA110 and AVA122, then sequenced with AVA037 using the Sanger method (Supplementary Table 2).

### Yeast fluorescent reporter assays

Dual fluorescent reporter plasmids containing an in-frame stop codon, a +1 frameshift, or a –1 frameshift were subjected to 5'-extended pegRNA or 3'-extended pegRNA prime editing reactions in vitro as described above using 100 ng of plasmid substrate. Following incubation at 37 °C for 1 h, the reactions were diluted with water and plasmid DNA was precipitated with 0.3 M sodium acetate and 70% ethanol. Resuspended DNA was transformed into *Saccharomyces cerevisiae* by electroporation as previously described<sup>42</sup> and plated on synthetic complete medium without leucine (SC(glucose), L–). GFP and mCherry fluorescence signals were visualized from colonies with the Typhoon FLA 7000 biomolecular imager.

### General mammalian cell culture conditions

HEK293T (ATCC CRL-3216), U2OS (ATCC HTB-96), K562 (CCL-243), and HeLa (CCL-2) cells were purchased from ATCC and cultured and passaged in Dulbecco's modified Eagle's medium (DMEM) plus GlutaMAX (ThermoFisher Scientific), McCoy's 5A medium (Gibco), RPMI medium 1640 plus GlutaMAX (Gibco), or Eagle's minimal essential medium (EMEM, ATCC), respectively, each supplemented with 10% (v/v) fetal bovine serum (Gibco, qualified) and 1× penicillin streptomycin (Corning). All cell types were incubated, maintained, and cultured at 37 °C with 5% CO<sub>2</sub>. Cell lines were authenticated by their respective suppliers and tested negative for mycoplasma.

### HEK293T tissue culture transfection protocol and genomic DNA preparation

HEK293T cells were seeded on 48-well poly-D-lysine coated plates (Corning). Between 16 and 24 h after seeding, cells were transfected at approximately 60% confluency with 1 μl lipofectamine 2000 (Thermo Fisher Scientific) according to the manufacturer's protocols and 750 ng PE plasmid, 250 ng pegRNA plasmid, and 83 ng sgRNA plasmid (for PE3 and PE3b). Unless otherwise stated, cells were cultured for 3 days following transfection, after which the medium was removed, the cells were washed with 1× PBS solution (Thermo Fisher Scientific), and genomic DNA was extracted by the addition of 150 μl of freshly prepared lysis buffer (10 mM Tris-HCl, pH 7.5; 0.05% SDS; 25 μg/ml proteinase K (ThermoFisher Scientific)) directly into each well of the tissue culture plate. The genomic DNA mixture was incubated at 37 °C for 1–2 h, followed by an 80 °C enzyme inactivation step for 30 min. Primers used for mammalian cell genomic DNA amplification are listed in Supplementary Table 4. For HDR experiments in HEK293T cells, 231 ng Cas9 nuclease-expression plasmid, 69 ng sgRNA-expression plasmid and 50 ng (1.51 pmol) of 100-nt ssDNA donor template (PAGE-purified; Integrated DNA Technologies) was lipofected using 1.4 μl lipofectamine 2000 (ThermoFisher) per well. Genomic DNA from all HDR experiments was purified using the Agencourt DNAAdvance Kit (Beckman Coulter), according to the manufacturer's protocol.

### High-throughput DNA sequencing of genomic DNA samples

Genomic sites of interest were amplified from genomic DNA samples and sequenced on an Illumina MiSeq as previously described with the following modifications<sup>17,18</sup>. In brief, amplification primers containing Illumina forward and reverse adapters (Supplementary Table 4) were used for a first round of PCR (PCR 1) to amplify the genomic region of interest. PCR 1 reactions (25 μl) were performed with 0.5 μM of each

forward and reverse primer, 1 µl genomic DNA extract and 12.5 µl Phusion U Green Multiplex PCR Master Mix. PCR reactions were carried out as follows: 98 °C for 2 min, then 30 cycles of [98 °C for 10 s, 61 °C for 20 s, and 72 °C for 30 s], followed by a final 72 °C extension for 2 min. Unique Illumina barcoding primer pairs were added to each sample in a secondary PCR reaction (PCR 2). Specifically, 25 µl of a given PCR 2 reaction contained 0.5 µM of each unique forward and reverse Illumina barcoding primer pair, 1 µl unpurified PCR 1 reaction mixture, and 12.5 µl Phusion U Green Multiplex PCR 2× Master Mix. The barcoding PCR 2 reactions were carried out as follows: 98 °C for 2 min, then 12 cycles of [98 °C for 10 s, 61 °C for 20 s, and 72 °C for 30 s], followed by a final 72 °C extension for 2 min. PCR products were evaluated analytically by electrophoresis in a 1.5% agarose gel. PCR 2 products (pooled by common amplicons) were purified by electrophoresis with a 1.5% agarose gel using a QIAquick Gel Extraction Kit (Qiagen), eluting with 40 µl water. DNA concentration was measured by fluorometric quantification (Qubit, ThermoFisher Scientific) or qPCR (KAPA Library Quantification Kit-Illumina, KAPA Biosystems) and sequenced on an Illumina MiSeq instrument according to the manufacturer's protocols.

Sequencing reads were demultiplexed using MiSeq Reporter (Illumina). Alignment of amplicon sequences to a reference sequence was performed using CRISPResso2<sup>43</sup>. For all prime editing yield quantification, prime editing efficiency was calculated as: percentage of (number of reads with the desired edit that do not contain indels)/(number of total reads). For quantification of point mutation editing, CRISPResso2 was run in standard mode with "discard\_indel\_reads" on. Prime editing for installation of point mutations was then explicitly calculated as: (frequency of specified point mutation in non-discarded reads) × (number of non-discarded reads)/(total reads). For insertion or deletion edits, CRISPResso2 was run in HDR mode using the desired allele as the expected allele (e flag), and with "discard\_indel\_reads" on. Editing yield was calculated as: (number of HDR-aligned reads)/(total reads). For all experiments, indel yields were calculated as: (number of indel-containing reads)/(total reads).

### Nucleofection of U2OS, K562, and HeLa cells

Nucleofection was used for transfection in all experiments using K562, HeLa, and U2OS cells. For PE conditions in these cell types, 800 ng prime editor expression plasmid, 200 ng pegRNA expression plasmid, and 83 ng nicking sgRNA expression plasmid was nucleofected in a final volume of 20 µl in a 16-well nucleocuvette strip (Lonza). For HDR conditions in these three cell types, 350 ng Cas9 nuclease expression plasmid, 150 ng sgRNA expression plasmid and 200 pmol (6.6 µg) 100-nt ssDNA donor template (PAGE-purified; Integrated DNA Technologies) was nucleofected in a final volume of 20 µl per sample in a 16-well Nucleocuvette strip (Lonza). K562 cells were nucleofected using the SF Cell Line 4D-Nucleofector X Kit (Lonza) with 5 × 10<sup>5</sup> cells per sample (program FF-120), according to the manufacturer's protocol. U2OS cells were nucleofected using the SE Cell Line 4D-Nucleofector X Kit (Lonza) with 3–4 × 10<sup>5</sup> cells per sample (program DN-100), according to the manufacturer's protocol. HeLa cells were nucleofected using the SE Cell Line 4D-Nucleofector X Kit (Lonza) with 2 × 10<sup>5</sup> cells per sample (program CN-114), according to the manufacturer's protocol. Cells were harvested 72 h after nucleofection for genomic DNA extraction.

### Genomic DNA extraction for HDR experiments

Genomic DNA from all HDR comparison experiments in HEK293T, HEK293T HBB(E6V), K562, U2OS, and HeLa cells was purified using the Agencourt DNAdvance Kit (Beckman Coulter), according to the manufacturer's protocol.

### Comparison between PE2, PE3, BE2, BE4max, ABEdmax, and ABEmax

HEK293T cells were seeded on 48-well poly-D-lysine coated plates (Corning). After 16–24 h, cells were transfected at approximately 60%

confluency. For base editing with CBE or ABE constructs, cells were transfected with 750 ng base editor plasmid, 250 ng sgRNA expression plasmid, and 1 µl of lipofectamine 2000 (Thermo Fisher Scientific). PE transfections were performed as described above. Genomic DNA extraction for PE and BE was performed as described above.

### Determination of PE3 activity at known Cas9 off-target sites

To evaluate PE3 off-target editing activity at known Cas9 off-target sites, genomic DNA extracted from HEK293T cells 3 days after transfection with PE3 was used as template for PCR amplification of 16 previously reported Cas9 off-target genomic sites<sup>32,33</sup> (the top four off-target sites each for the *HEK3*, *EMX1*, *FANCF*, and *HEK4* spacers; primer sequences are listed in Supplementary Table 4). These genomic DNA samples were identical to those used for quantifying on-target PE3 editing activities shown in Fig. 4 or Extended Data Fig. 5d, e; pegRNA and nicking sgRNA sequences are listed in Supplementary Table 3. Following PCR amplification of off-target sites, amplicons were sequenced on the Illumina MiSeq platform as described above (see 'High-throughput DNA sequencing of genomic DNA samples' section). To determine the on-target and off-target editing activity of Cas9 nuclease, Cas9(H840A) nickase, dCas9, and PE2-dRT, we transfected HEK293T cells with 750 ng editor plasmid (Cas9 nuclease, Cas9(H840A) nickase, dCas9, or PE2-dRT), 250 ng pegRNA or sgRNA plasmid, and 1 µl lipofectamine 2000. Genomic DNA was isolated from cells 3 days after transfection as described above. On-target and off-target genomic loci were amplified by PCR using the primer sequences in Supplementary Table 4 and sequenced on an Illumina MiSeq.

High-throughput sequencing (HTS) data analysis was performed using CRISPResso2<sup>43</sup>. The editing efficiencies of Cas9 nuclease, Cas9 H840A nickase, and dCas9 were quantified as the percentage of total sequencing reads containing indels. For quantification of PE3 and PE3-dRT off-targets, aligned sequencing reads were examined for point mutations, insertions, or deletions that were consistent with the anticipated product of pegRNA reverse transcription initiated at the Cas9 nick site. Single nucleotide variations occurring at <0.1% overall frequency among total reads within a sample were excluded from analysis. For reads containing single nucleotide variations that both occurred at frequencies ≥0.1% and were partially consistent with the pegRNA-encoded edit, *t*-tests (unpaired, one-tailed,  $\alpha=0.5$ ) were used to determine whether the variants occurred at significantly higher levels compared to samples treated with pegRNAs that contained the same spacer but encoded different edits. To avoid differences in sequencing errors, comparisons were made between samples that were sequenced simultaneously within the same MiSeq run. Variants that did not meet the criteria of  $P > 0.05$  were excluded. Off-target PE3 editing activity was then calculated as the percentage of total sequencing reads that met the above criteria.

### Generation of a HEK293T cell line containing HBB(E6V) using Cas9-initiated HDR

HEK293T cells were seeded in a 48-well plate and transfected at approximately 60% confluency with 1.5 µl lipofectamine 2000, 300 ng Cas9(D10A) nickase plasmid, 100 ng sgRNA plasmid, and 200 ng 100-mer ssDNA donor template (Supplementary Table 5). Three days after transfection, the medium was exchanged for fresh medium. Four days after transfection, cells were dissociated using 30 µl TrypLE solution and suspended in 1.5 ml medium. Single cells were isolated into individual wells of two 96-well plates by fluorescence-activated cell sorting (FACS) (Beckman-Coulter Astrios). See Supplementary Note 1 for representative FACS sorting examples. Cells were expanded for 14 days before genomic DNA sequencing as described above. Of the isolated clonal populations, none was found to be homozygous for the *HBB* allele encoding the E6V mutation, so a second round of editing by lipofection, sorting, and outgrowth was repeated in a partially edited cell line to yield a cell line homozygous for the E6V-encoding allele.



# Article

## Generation of a HEK293T cell line containing HBB(E6V) using PE3

HEK293T cells ( $2.5 \times 10^4$ ) were seeded on 48-well poly-D-lysine coated plates (Corning). Between 16 and 24 h after seeding, cells were transfected at approximately 70% confluency with 1  $\mu$ l lipofectamine 2000 (Thermo Fisher Scientific) according to the manufacturer's protocols and 750 ng PE2-P2A-GFP plasmid, 250 ng pegRNA plasmid, and 83 ng sgRNA plasmid. After 3 days, cells were washed with  $1 \times$  PBS (Gibco) and dissociated using TrypLE Express (Gibco). Cells were then diluted with DMEM plus GlutaMax (Thermo Fisher Scientific) supplemented with 10% (v/v) FBS (Gibco) and passed through a 35- $\mu$ m cell strainer (Corning) before sorting. Flow cytometry was carried out on a LE-MA900 cell sorter (Sony). Cells were treated with 3 nM DAPI (BioLegend) 15 min before sorting. After gating for doublet exclusion, single DAPI-negative cells with GFP fluorescence above that of a GFP-negative control cell population were sorted into 96-well flat-bottom cell culture plates (Corning) filled with pre-chilled DMEM with GlutaMax supplemented with 10% FBS. See Supplementary Note 1 for representative FACS sorting examples and allele tables. Cells were cultured for 10 days before genomic DNA extraction and characterization by HTS, as described above. A total of six clonal cell lines were identified that are homozygous for the E6V-encoding mutation in *HBB*.

## Generation of a HEK293T cell line containing the *HEXA*<sup>1278+TATC</sup> insertion using PE3

HEK293T cells containing the *HEXA*<sup>1278+TATC</sup> allele were generated following the protocol described above for creation of the HBB(E6V) cell line; pegRNA and sgRNA sequences are listed in Supplementary Table 3 under the Fig. 5 subheading. After transfection and sorting, cells were cultured for 10 days before genomic DNA was extracted and characterized by HTS, as described above. We recovered two heterozygous cell lines that contained 50% *HEXA*<sup>1278+TATC</sup> alleles and two homozygous cell lines containing 100% *HEXA*<sup>1278+TATC</sup> alleles.

## Cell viability assays

HEK293T cells were seeded in 48-well plates and transfected at approximately 70% confluency with 750 ng editor plasmid (PE2, PE2(R110S/K103L), Cas9(H840A) nickase, or dCas9), 250 ng HEK3-targeting pegRNA plasmid, and 1  $\mu$ l lipofectamine 2000, as described above. Cell viability was measured every 24 h post-transfection for 3 days using the CellTiter-Glo 2.0 assay (Promega) according to the manufacturer's protocol. Luminescence was measured in 96-well flat-bottomed polystyrene microplates (Corning) using a M1000 Pro microplate reader (Tecan) with a 1-s integration time.

## Lentivirus production

Lentivirus was produced as previously described<sup>44</sup>. T-75 flasks of rapidly dividing HEK293T cells (ATCC; Manassas, VA, USA) were transfected with lentivirus production helper plasmids pVSV-G and psPAX2 in combination with modified lentiCRISPRv2 genomes carrying intein-split PE2 editor using FuGENE HD (Promega, Madison, WI, USA) according to the manufacturer's protocol. Four split-intein editor constructs were designed: 1) a viral genome encoding a U6-pegRNA expression cassette and the N-terminal portion (1–573) of Cas9(H840A) nickase fused to the Npu N-intein, a self-cleaving P2A peptide, and GFP-KASH; 2) a viral genome encoding the Npu C-intein fused to the C-terminal remainder of PE2; 3) a viral genome encoding the Npu C-intein fused to the C-terminal remainder of Cas9 for the Cas9 control; and 4) a nicking sgRNA for *DNMT1* (derived from Addgene plasmid no. 52963). The split-intein<sup>45</sup> mediates *trans* splicing to join the two halves of PE2 or Cas9, while the P2A GFP-KASH enables co-translational production of a nuclear membrane-localized GFP. After 48 h, supernatant was collected, centrifuged at 500g for 5 min to remove cellular debris, and filtered using a 0.45- $\mu$ m filter. Filtered supernatant was concentrated using the PEG-it Virus Precipitation Solution (System Biosciences, Palo Alto, CA, USA) according to the manufacturer's directions. The resulting pellet

was resuspended in Opti-MEM (Thermo Fisher Scientific, Waltham, MA, USA) using 1% of the original medium volume. Resuspended pellet was flash-frozen and stored at  $-80^\circ\text{C}$  until use.

## Mouse primary cortical neuron dissection and culture

E18.5 dissociated cortical cultures were taken from timed-pregnant C57BL/6 mice (Charles River). Embryos were removed from pregnant mice after euthanasia by  $\text{CO}_2$  followed by decapitation. Cortical caps were dissected in ice-cold Hibernate-E supplemented with penicillin/streptomycin (Life Technologies). Following a rinse with ice-cold Hibernate-E, tissue was digested at  $37^\circ\text{C}$  for 8 min in papain/DNase (Worthington/Sigma). Tissue was triturated in NBActiv4 (BrainBits) supplemented with DNase. Cells were counted and plated in 24-well plates at 100,000 cells per well. Half of the medium was changed twice per week.

## Prime editing in primary neurons and nucleus isolation

At days in vitro (DIV) 1, 15  $\mu$ l lentivirus was added at a 10:10:1 ratio of N-terminal:C-terminal:nicking sgRNA. At DIV 14, neuronal nuclei were isolated using the EZ-PREP buffer (Sigma D8938) following the manufacturer's protocol. All steps were performed on ice or at  $4^\circ\text{C}$ . Medium was removed from dissociated cultures, and cultures were washed with ice-cold PBS. PBS was aspirated and replaced with 200  $\mu$ l EZ-PREP solution. Following a 5-min incubation on ice, EZ-PREP was pipetted across the surface of the well to dislodge remaining cells. The sample was centrifuged at 500g for 5 min, and the supernatant removed. Samples were washed with 200  $\mu$ l EZ-PREP and centrifuged again at 500g for 5 min. Samples were resuspended with gentle pipetting in 200  $\mu$ l ice-cold Nuclei Suspension Buffer (NSB) consisting of 100  $\mu$ g/ml BSA and 3.33  $\mu$ M Vybrant DyeCycle Ruby (Thermo Fisher) in  $1 \times$  PBS, then centrifuged at 500g for 5 min. The supernatant was removed and nuclei were resuspended in 100  $\mu$ l NSB and sorted into 100  $\mu$ l Agencourt DNAdvance lysis buffer using a MoFlo Astris (Beckman Coulter) at the Broad Institute flow cytometry facility. Genomic DNA was purified according to the manufacturer's Agencourt DNAdvance instructions.

## RNA-seq and data analysis

HEK293T cells were co-transfected with *PRNP*-targeting or *HEXA*-targeting pegRNAs and PE2, PE2-dRT, or Cas9(H840A) nickase. Seventy-two hours after transfection, total RNA was harvested from cells using TRIzol reagent (Thermo Fisher) and purified with RNeasy Mini kit (Qiagen) including on-column DNaseI treatment. Ribosomes were depleted from total RNA using the rRNA removal protocol of the TruSeq Stranded Total RNA library prep kit (Illumina) and subsequently washed with RNAClean XP beads (Beckman Coulter). Sequencing libraries were prepared using ribo-depleted RNA on a SMARTer PrepX Apollo NGS library prep system (Takara) following the manufacturer's protocol. The resulting libraries were visualized on a 2200 TapeStation (Agilent Technologies), normalized using a Qubit dsDNA HS assay (Thermo Fisher), and sequenced on a NextSeq 550 using high output v2 flow cell (Illumina) as 75-bp paired-end reads. Fastq files were generated with bcl2fastq2 version 2.20 and trimmed using TrimGalore version 0.6.2 (<https://github.com/FelixKrueger/TrimGalore>) to remove low-quality bases, unpaired sequences, and adaptor sequences. Trimmed reads were aligned to a *Homo sapiens* genome assembly GRCh38 with a custom Cas9(H840A) gene entry using RSEM version 1.3.1<sup>46</sup>. The limma-voom<sup>47</sup> package was used to normalize gene expression levels and perform differential expression analysis with batch effect correction. Differentially expressed genes were called with FDR-corrected  $P < 0.05$  and fold change  $> 2$  cutoffs, and results were visualized in R.

## ClinVar analysis

The ClinVar variant summary was downloaded from NCBI (accessed July 15, 2019), and the information contained therein was used for all downstream analysis. The list of all reported variants was filtered by allele ID in order to remove duplicates and by clinical significance in order

to restrict the analysis to pathogenic variants. The list of pathogenic variants was filtered sequentially by variant type in order to calculate the fraction of pathogenic variants that are insertions, deletions, and so on. Single nucleotide variants (SNVs) were separated into two categories (transitions and transversions) on the basis of the reported reference and alternate alleles. SNVs that did not report reference or alternate alleles were excluded from the analysis.

The lengths of reported insertions, deletions, and duplications were calculated using reference/alternate alleles, variant start/stop positions, or appropriate identifying information in the variant name. Variants that did not report any of the above information were excluded from the analysis. The lengths of reported indels (single variants that include both insertions and deletions relative to the reference genome) were calculated by determining the number of mismatches or gaps in the best pairwise alignment between the reference and alternate alleles. Frequency distributions of variant lengths were calculated using GraphPad Prism 8.

### Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

### Data availability

High-throughput sequencing data have been deposited to the NCBI Sequence Read Archive database under accession PRJNA565979. Plasmids encoding PE1, PE2 (same as PE3), and pegRNA expression vectors are available from Addgene. Previously described plasmids expressing sgRNAs are also available from Addgene, such as Addgene plasmid no. 65777.

### Code availability

The script used to quantify pegRNA scaffold insertion is provided as Supplementary Note 4.

39. Anzalone, A. V., Lin, A. J., Zairis, S., Rabadan, R. & Cornish, V. W. Reprogramming eukaryotic translation with ligand-responsive synthetic RNA switches. *Nat. Methods* **13**, 453–458 (2016).
40. Badran, A. H. et al. Continuous evolution of *Bacillus thuringiensis* toxins overcomes insect resistance. *Nature* **533**, 58–63 (2016).
41. Anders, C. & Jinek, M. in *Methods in Enzymology* (eds. Doudna, J. A. & Sontheimer, E. J.) **546**, 1–20 (Academic, 2014).
42. Pirakitikulr, N., Ostrov, N., Peralta-Yahya, P. & Cornish, V. W. PCRless library mutagenesis via oligonucleotide recombination in yeast. *Protein Sci.* **19**, 2336–2346 (2010).
43. Clement, K. et al. CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat. Biotechnol.* **37**, 224–226 (2019).
44. Levy, J. M. & Nicoll, R. A. Membrane-associated guanylate kinase dynamics reveal regional and developmental specificity of synapse stability. *J. Physiol. (Lond.)* **595**, 1699–1709 (2017).
45. Zettler, J., Schütz, V. & Mootz, H. D. The naturally split Npu DnaE intein exhibits an extraordinarily high rate in the protein trans-splicing reaction. *FEBS Lett.* **583**, 909–914 (2009).
46. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
47. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

**Acknowledgements** We thank J. M. Madison for neuron cell culture advice. This work was supported by the Merkin Institute of Transformative Technologies in Healthcare, US NIH grants U01AI142756, RM1HG009490, R01EB022376, and R35GM118062, and the HHMI. A.V.A. acknowledges a Jane Coffin Childs postdoctoral fellowship. P.B.R. and A.R. acknowledge NIH T32 GM095450. A.A.S. acknowledges NIH T32 GM007726. P.J.C. and A.R. acknowledge NSF graduate fellowships. C.W. acknowledges a Damon Runyon Cancer Research Foundation fellowship (DRG-2343-18). G.A.N. acknowledges a Helen Hay Whitney postdoctoral fellowship.

**Author contributions** A.V.A. designed the research, performed experiments, analysed data, and wrote the manuscript. P.B.R., J.R.D., A.A.S., and G.A.N. performed human cell experiments and analysed data. L.W.K. and J.M.L. performed neuron experiments. P.J.C. and C.W. performed and analysed RNA-seq experiments. A.R. analysed ClinVar data. D.R.L. designed and supervised the research and wrote the manuscript.

**Competing interests** Authors through the Broad Institute have filed patent applications on prime editing. D.R.L. is a consultant and co-founder of Prime Medicine, Beam Therapeutics, Pairwise Plants, and Editas Medicine, companies that use genome editing.

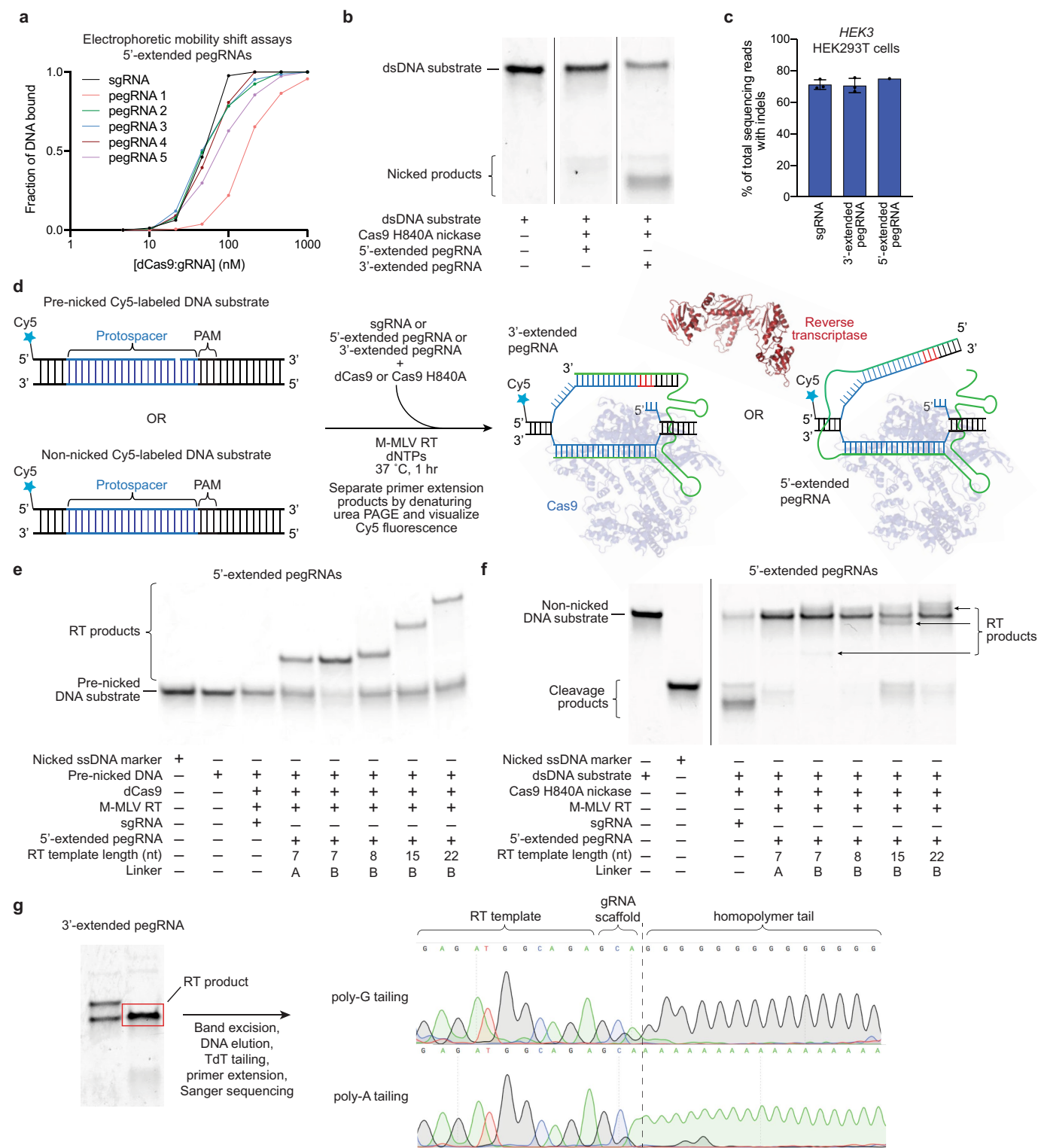
### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1711-4>.

**Correspondence and requests for materials** should be addressed to D.R.L.

**Peer review information** Nature thanks Guangping Gao, Randall Platt and Fyodor Urnov for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

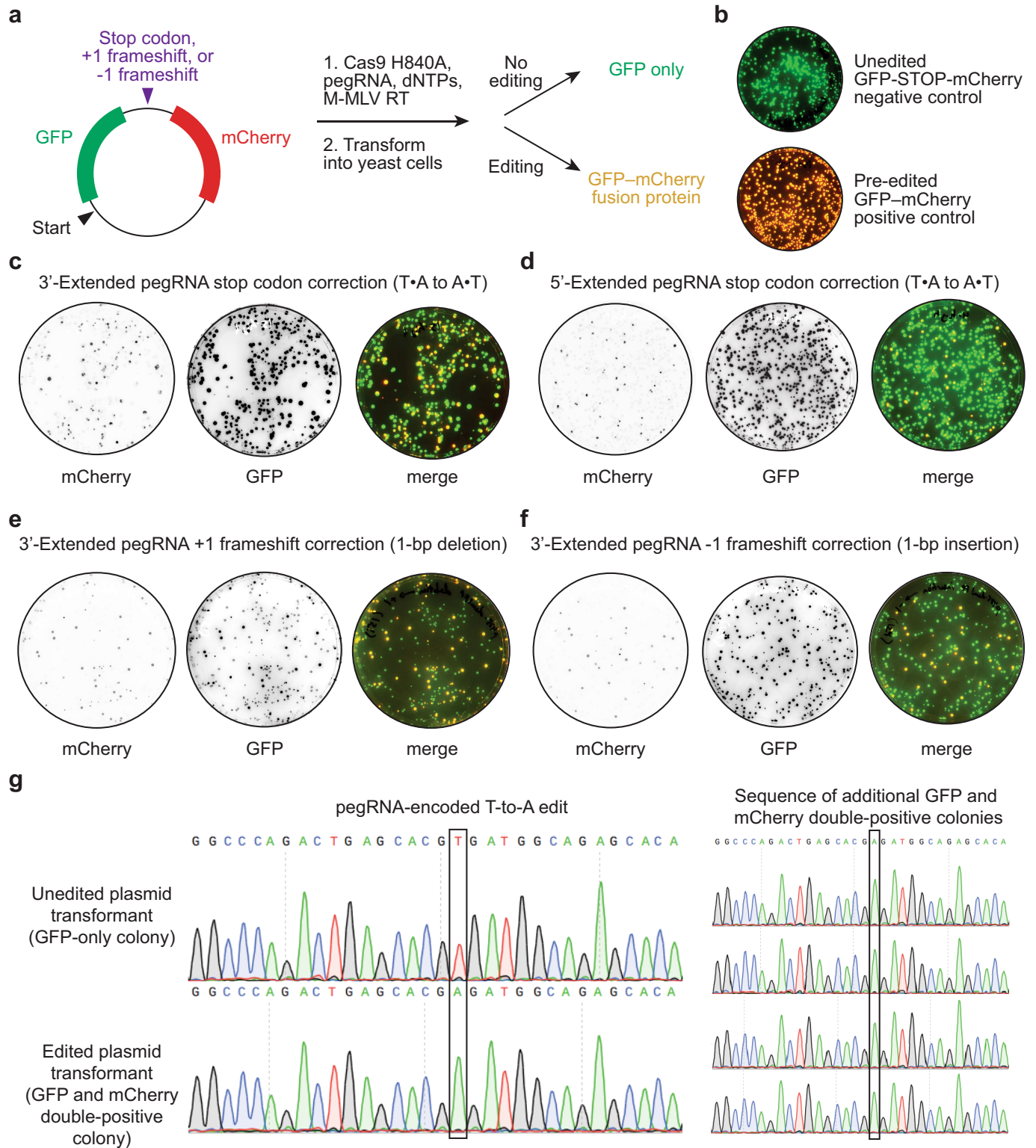


Extended Data Fig. 1| See next page for caption.

**Extended Data Fig. 1 | In vitro prime editing validation studies with**

**fluorescently labelled DNA substrates. a,** Electrophoretic mobility shift assays with dCas9, 5'-extended pegRNAs and 5'-Cy5-labelled DNA substrates. pegRNAs 1–5 contain a 15-nt linker sequence (linker A for pegRNA 1, linker B for pegRNAs 2–5) between the spacer and the PBS, a 5-nt PBS sequence, and RT templates of 7 nt (pegRNAs 1 and 2), 8 nt (pegRNA 3), 15 nt (pegRNA 4), and 22 nt (pegRNA 5). pegRNAs are those used in **e** and **f**; full sequences are listed in Supplementary Table 2. **b,** In vitro nicking assays of Cas9(H840A) using 5'-extended and 3'-extended pegRNAs. Data in **a**, **b** are representative of  $n = 2$  independent replicates. **c,** Cas9-mediated indel formation in HEK293T cells at *HEK3* using 5'-extended and 3'-extended pegRNAs. Mean  $\pm$  s.d. of  $n = 3$  independent biological replicates. **d,** Overview of prime editing in vitro biochemical assays. 5'-Cy5-labelled pre-nicked and non-nicked dsDNA substrates were tested. sgRNAs, 5'-extended pegRNAs, or 3'-extended pegRNAs were pre-complexed with dCas9 or Cas9(H840A) nickase, then combined with dsDNA substrate, Superscript III M-MLV RT, and dNTPs. Reactions were allowed to proceed at 37 °C for 1 h before separation by denaturing urea PAGE and visualization by Cy5 fluorescence. **e,** Primer

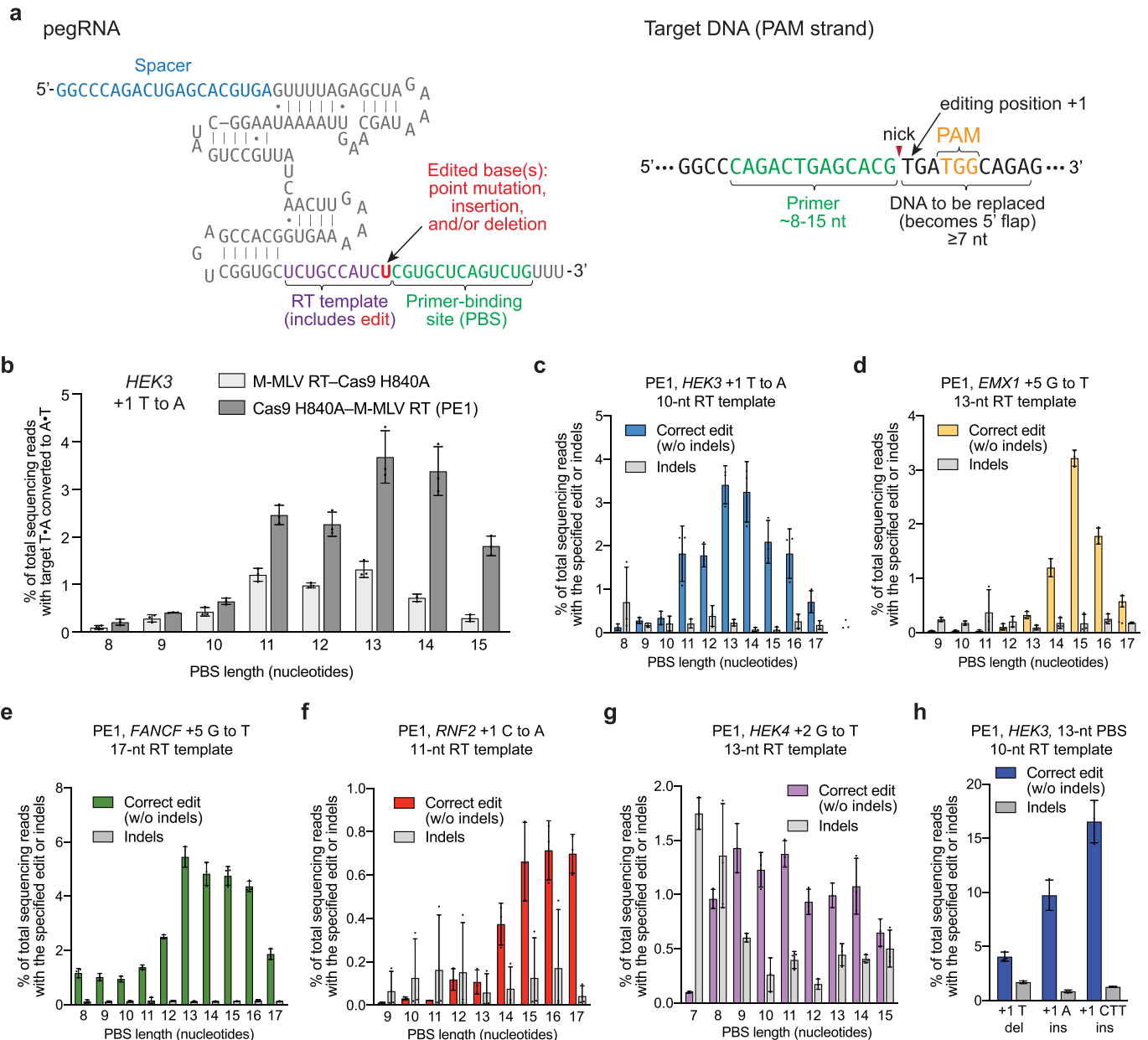
extension reactions using 5'-extended pegRNAs, pre-nicked DNA substrates, and dCas9 lead to substantial conversion to RT products. **f,** Primer extension reactions using 5'-extended pegRNAs as in **b** with non-nicked DNA substrate and Cas9(H840A) nickase. Product yields are greatly reduced by comparison to pre-nicked substrate. **g,** An in vitro primer extension reaction using a 3'-pegRNA generates a single apparent product by denaturing urea PAGE. The RT product band was excised, eluted from the gel, then subjected to homopolymer tailing with terminal transferase (TdT) using either dGTP or dATP. Tailed products were extended using poly-T or poly-C primers, and the resulting DNA was sequenced. Sanger traces indicate that three nucleotides derived from the pegRNA scaffold were reverse-transcribed (added as the final 3' nucleotides to the DNA product). Note that pegRNA scaffold insertion is much rarer in mammalian cell prime editing experiments than in vitro (Extended Data Fig. 6), potentially owing to the inability of the tethered RT to access the Cas9-bound guide RNA scaffold, and/or cellular excision of mismatched 3' ends of 3' flaps containing pegRNA scaffold sequences. Data in **e–g** are representative of  $n = 2$  independent replicates. For gel source data, see Supplementary Fig. 1.



**Extended Data Fig. 2 | Cellular repair in yeast of 3' DNA flaps from in vitro prime editing reactions.** **a**, Dual fluorescent protein reporter plasmids contain GFP and mCherry open reading frames separated by a target site encoding an in-frame stop codon, a +1 frameshift, or a -1 frameshift. Prime editing reactions were carried out in vitro with Cas9(H840A) nickase, pegRNA, dNTPs, and M-MLV RT, then transformed into yeast. Colonies that contain unedited plasmids produce GFP but not mCherry. Yeast colonies containing edited plasmids produce both GFP and mCherry as a fusion protein. **b**, Overlay of GFP and mCherry fluorescence for yeast colonies transformed with reporter plasmids containing a stop codon between GFP and mCherry (unedited

negative control, top), or containing no stop codon or frameshift between GFP and mCherry (pre-edited positive control, bottom). **c-f**, Visualization of mCherry and GFP fluorescence from yeast colonies transformed with in vitro prime editing reaction products. **c, d**, Stop codon correction via T•A-to-A•T transversion using a 3'-extended pegRNA (**c**) or a 5'-extended pegRNA (**d**). **e**, +1 frameshift correction via a 1-bp deletion using a 3'-extended pegRNA. **f**, -1 frameshift correction via a 1-bp insertion using a 3'-extended pegRNA. **g**, Sanger DNA sequencing traces from plasmids isolated from GFP-only colonies in **b** and GFP and mCherry double-positive colonies in **c**. Data in **b-g** are representative of  $n = 2$  independent replicates.

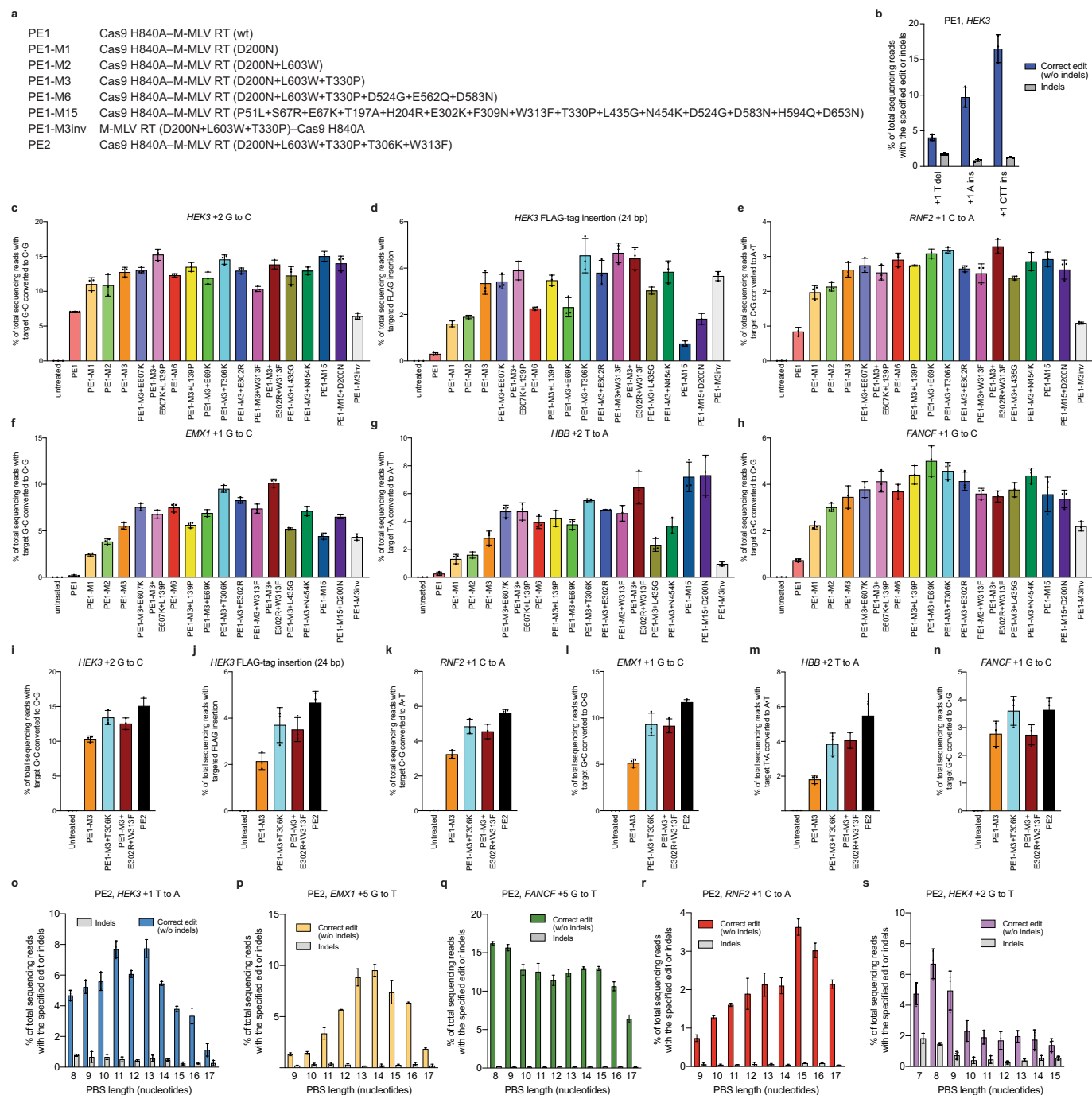




**Extended Data Fig. 3 | Prime editing of genomic DNA in human cells by PE1.**

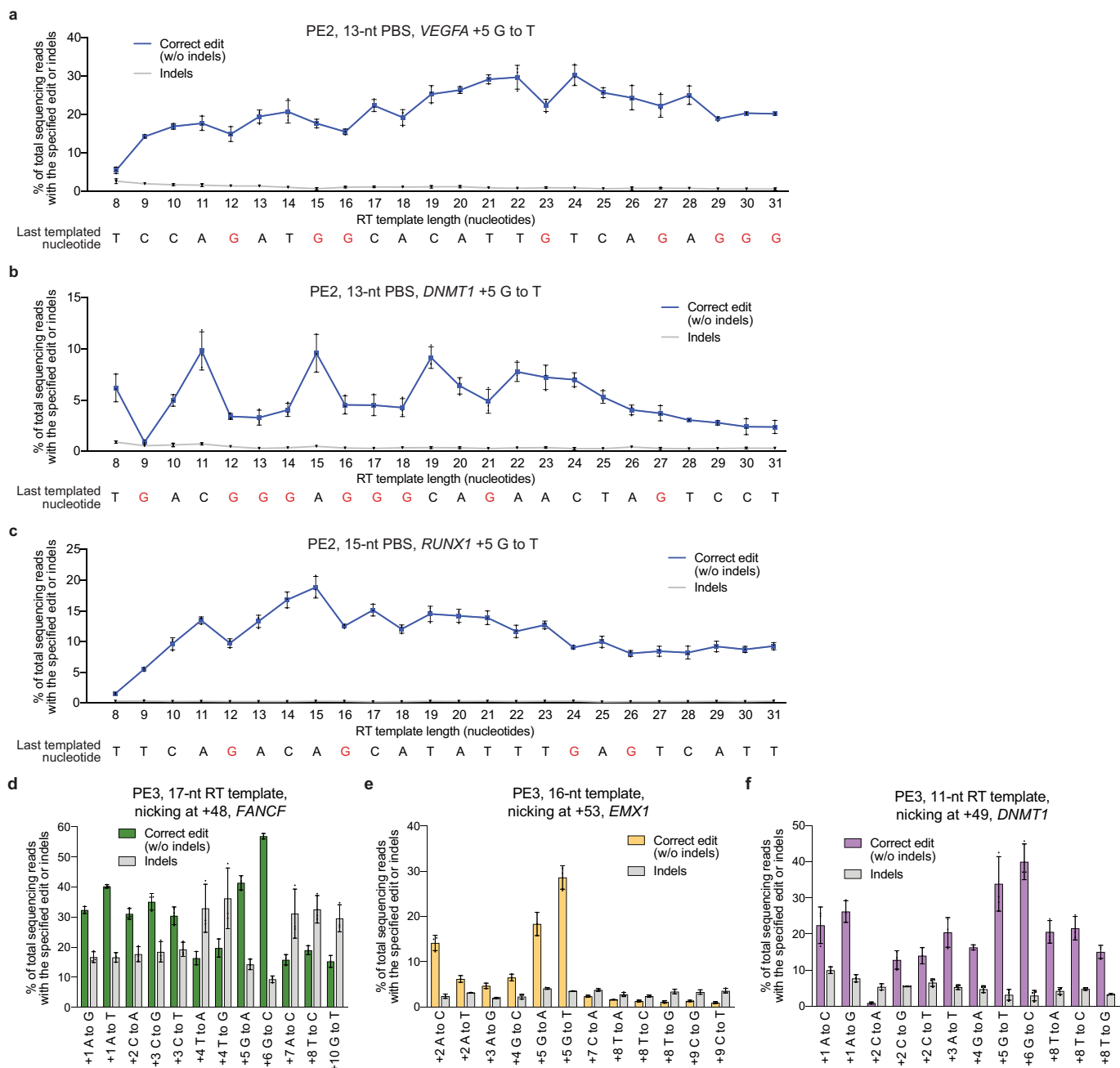
**a**, pegRNAs contain a spacer sequence, an sgRNA scaffold, and a 3' extension containing an RT template (purple), which contains the edited base(s) (red), and a primer-binding site (PBS, green). The primer-binding site hybridizes to the nicked target DNA strand. The RT template is homologous to the DNA sequence downstream of the nick, with the exception of the encoded edited base(s). **b**, Installation of a T→A-to-A→T transversion at the *HEK3* site in HEK293T cells using Cas9(H840A) nickase fused to wild-type M-MLV RT (PE1) and pegRNAs with varying PBS lengths. **c**, T→A-to-A→T transversion editing efficiency and indel generation by PE1 at the +1 position of *HEK3* using pegRNAs containing 10-nt RT templates and PBS sequences ranging from 8 to 17 nt. **d**, G→C-to-T→A transversion editing efficiency and indel generation by PE1 at the +5 position of *EMX1* using pegRNAs containing 13-nt RT templates and PBS

sequences ranging from 9 to 17 nt. **e**, G→C-to-T→A transversion editing efficiency and indel generation by PE1 at the +5 position of *FANCF* using pegRNAs containing 17-nt RT templates and PBS sequences ranging from 8 to 17 nt. **f**, C→G-to-A→T transversion editing efficiency and indel generation by PE1 at the +1 position of *RNF2* using pegRNAs containing 11-nt RT templates and PBS sequences ranging from 9 to 17 nt. **g**, G→C-to-T→A transversion editing efficiency and indel generation by PE1 at the +2 position of *HEK4* using pegRNAs containing 13-nt RT templates and PBS sequences ranging from 7 to 15 nt. **h**, PE1-mediated +1 T deletion, +1 A insertion, and +1 CTT insertion at the *HEK3* site using a 13-nt PBS and a 10-nt RT template. Sequences of pegRNAs are as in Fig. 2a (Supplementary Table 3). Editing efficiencies reflect sequencing reads that contain the intended edit and do not contain indels among all treated cells, with no sorting. Mean ± s.d. of  $n = 3$  independent biological replicates.



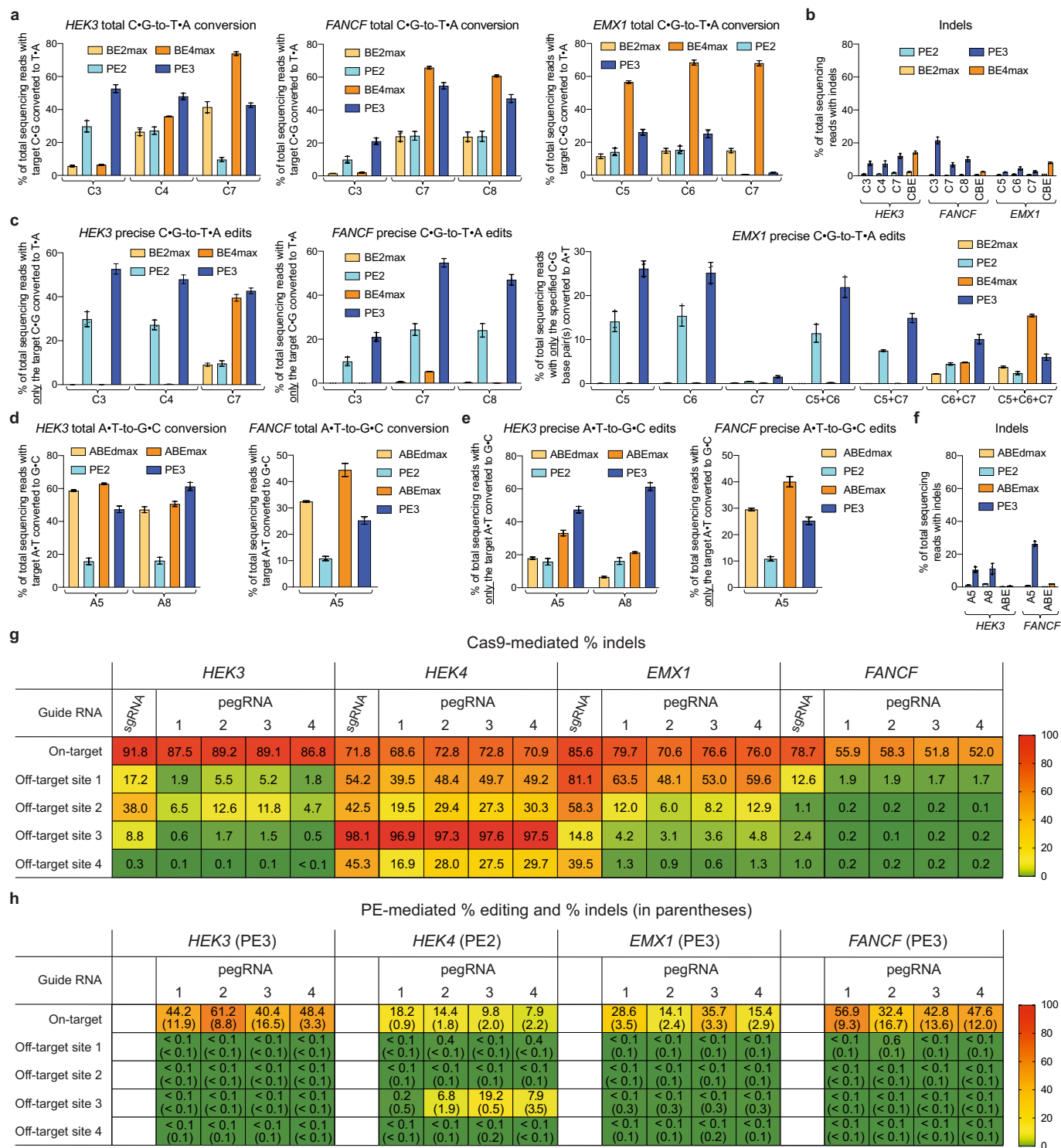
**Extended Data Fig. 4 | Evaluation of M-MLV RT variants for prime editing. a,** Abbreviations for prime editor variants used in this figure. **b,** Targeted insertion and deletion edits with PE1 at the *HEK3* locus. **c-h,** Comparison of 18 prime editor constructs containing M-MLV RT variants for their ability to install a +2 G-C to C-G transversion edit at *HEK3* (c), a 24-bp Flag insertion at the +1 position of *HEK3* (d), a +1 C-G to A-T transversion edit at *RNF2* (e), a +1 G-C to C-G transversion edit at *EMX1* (f), a +2 T-A to A-T transversion edit at *HBB* (g), and a +1 G-C to C-G transversion edit at *FANCF* (h). **i-n,** Comparison of four

prime editor constructs containing M-MLV variants for their ability to install the edits shown in c-h in a second round of independent experiments. **o-s,** PE2 editing efficiency at five genomic loci with varying PBS lengths. **o,** +1 T-A to A-T at *HEK3*. **p,** +5 G-C to T-A at *EMX1*. **q,** +5 G-C to T-A at *FANCF*. **r,** +1 C-G to A-T at *RNF2*. **s,** +2 G-C to T-A at *HEK4*. Editing efficiencies reflect sequencing reads that contain the intended edit and do not contain indels among all treated cells, with no sorting. Mean  $\pm$  s.d. of  $n = 3$  independent biological replicates.



**Extended Data Fig. 5 | Design features of pegRNA PBS and RT template sequences, and additional editing examples with PE3. a**, PE2-mediated +5 G•C-to-T•A transversion editing efficiency (blue line) at *VEGFA* in HEK293T cells as a function of RT template length. Indels (grey line) are plotted for comparison. The sequence below the graph shows the last nucleotide templated for synthesis by the pegRNA. G nucleotides (templated by a C in the

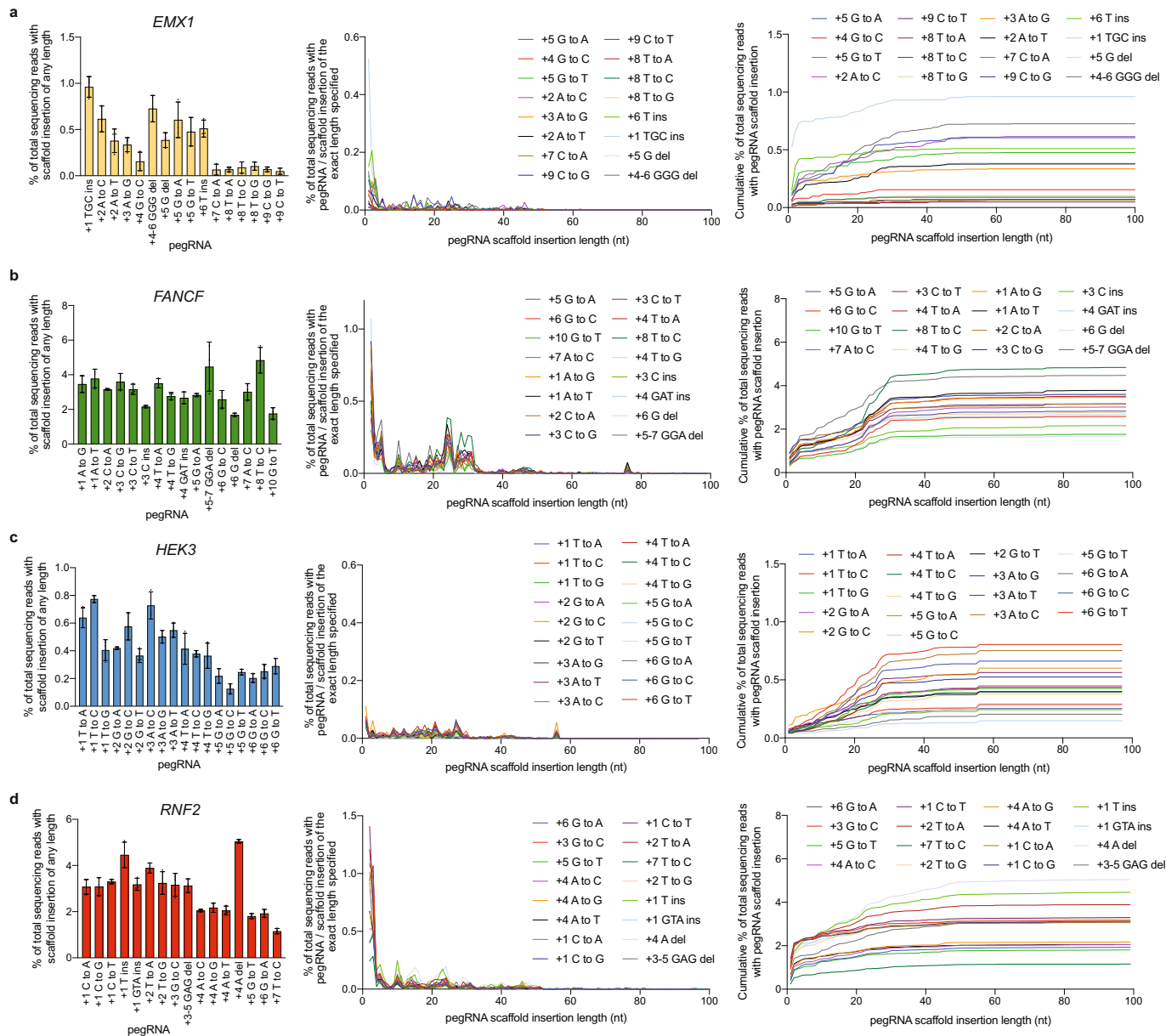
pegRNA) are highlighted in red; RT templates that end in C should be avoided during pegRNA design to maximize prime editing efficiency. **b**, +5 G•C-to-T•A transversion editing and indels for *DNMT1* as in **a**. **c**, +5 G•C-to-T•A transversion editing and indels for *RUNX1* as in **a**. **d-f**, PE3-mediated transition and transversion edits at the specified positions for *FANCF* (**d**), *EMX1* (**e**), and *DNMT1* (**f**). Mean  $\pm$  s.d. of  $n=3$  independent biological replicates.



**Extended Data Fig. 6 | Comparison of prime editing and base editing, and off-target editing by Cas9 and prime editors at known Cas9 off-target sites.**

**a**, C•G-to-T•A editing efficiency at the same target nucleotides for PE2, PE3, BE2max, and BE4max at endogenous *HEK3*, *FANCF*, and *EMX1* sites in HEK293T cells. **b**, Indel frequency from treatments in **a**. **c**, Editing efficiency of precise C•G-to-T•A edits (without bystander edits or indels) at *HEK3*, *FANCF*, and *EMX1*. **d**, Total A•T-to-G•C editing efficiency for PE2, PE3, ABE2max, and ABE4max at *HEK3* and *FANCF*. **e**, Precise A•T-to-G•C editing efficiency without bystander edits or indels at *HEK3* and *FANCF*. **f**, Indel frequency from treatments in **d**. **g**, Average triplicate Cas9 nuclease editing efficiencies (indel frequencies) in

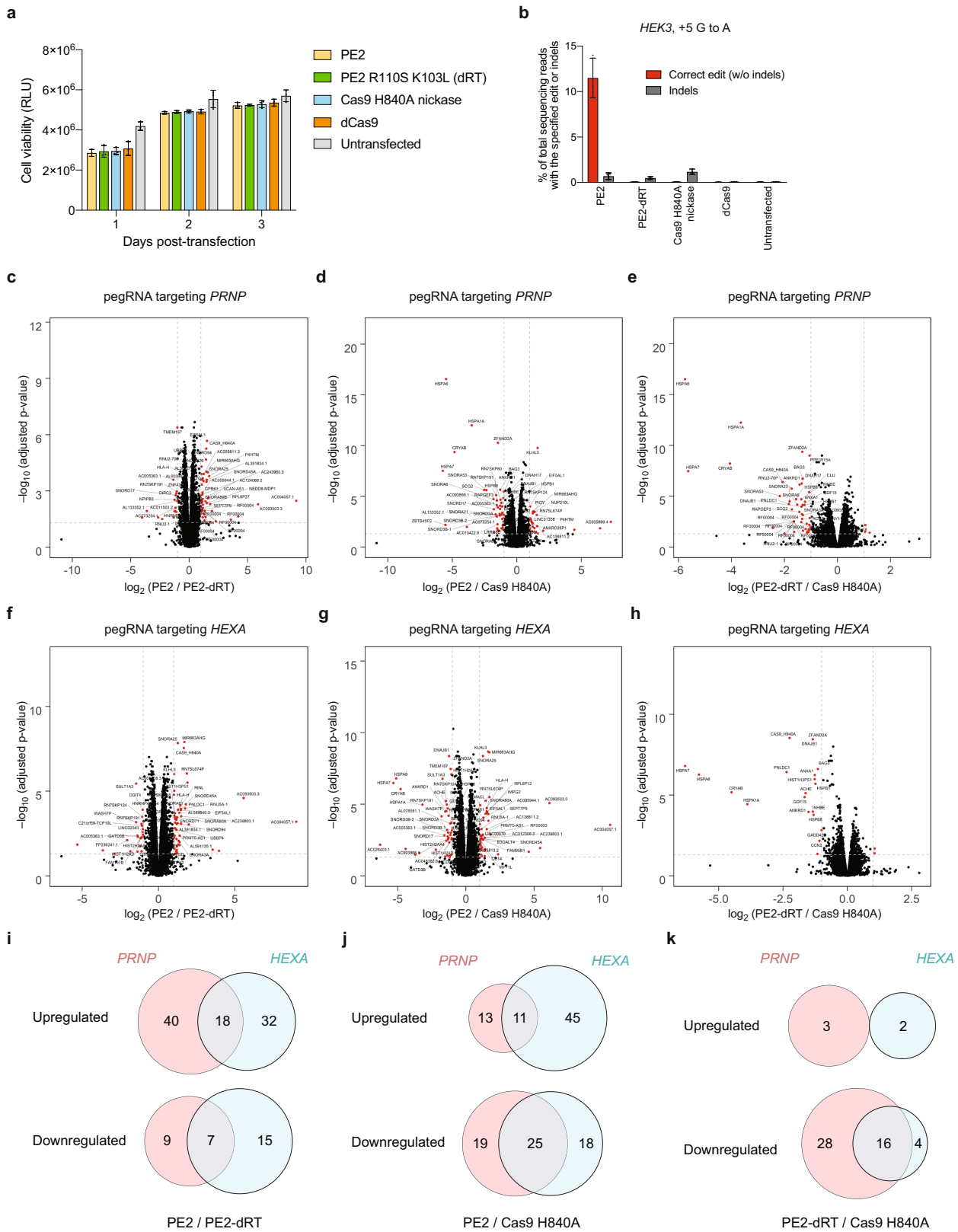
HEK293T cells at four endogenous on-target sites and their 16 known top off-target sites<sup>32,33</sup>. For each on-target site, Cas9 was paired with an sgRNA or with each of four pegRNAs that recognize the same protospacer. **h**, Average triplicate on-target and off-target editing efficiencies and indel efficiencies (below in parentheses) in HEK293T cells for PE2 or PE3 paired with each pegRNA in **g**. Editing efficiencies reflect sequencing reads that contain the intended edit and do not contain indels among all treated cells, with no sorting. Off-target editing efficiencies in **h** reflect off-target locus modification consistent with prime editing. Mean  $\pm$  s.d. of  $n = 3$  independent biological replicates.



**Extended Data Fig. 7 | Incorporation of pegRNA scaffold sequence into target loci.** HTS data were analysed for pegRNA scaffold sequence insertion as described in Supplementary Note 4. **a**, Analysis for the *EMX1* locus. Shown is the percentage of total sequencing reads containing one or more pegRNA scaffold sequence nucleotides within an insertion adjacent to the RT template (left); the

percentage of total sequencing reads containing a pegRNA scaffold sequence insertion of the specified length (middle); and the cumulative total percentage of pegRNA insertion up to and including the length specified on the x-axis. **b**, As in **a** for *FANCF*. **c**, As in **a** for *HEK3*. **d**, As in **a** for *RNF2*. Mean  $\pm$  s.d. of  $n=3$  independent biological replicates.



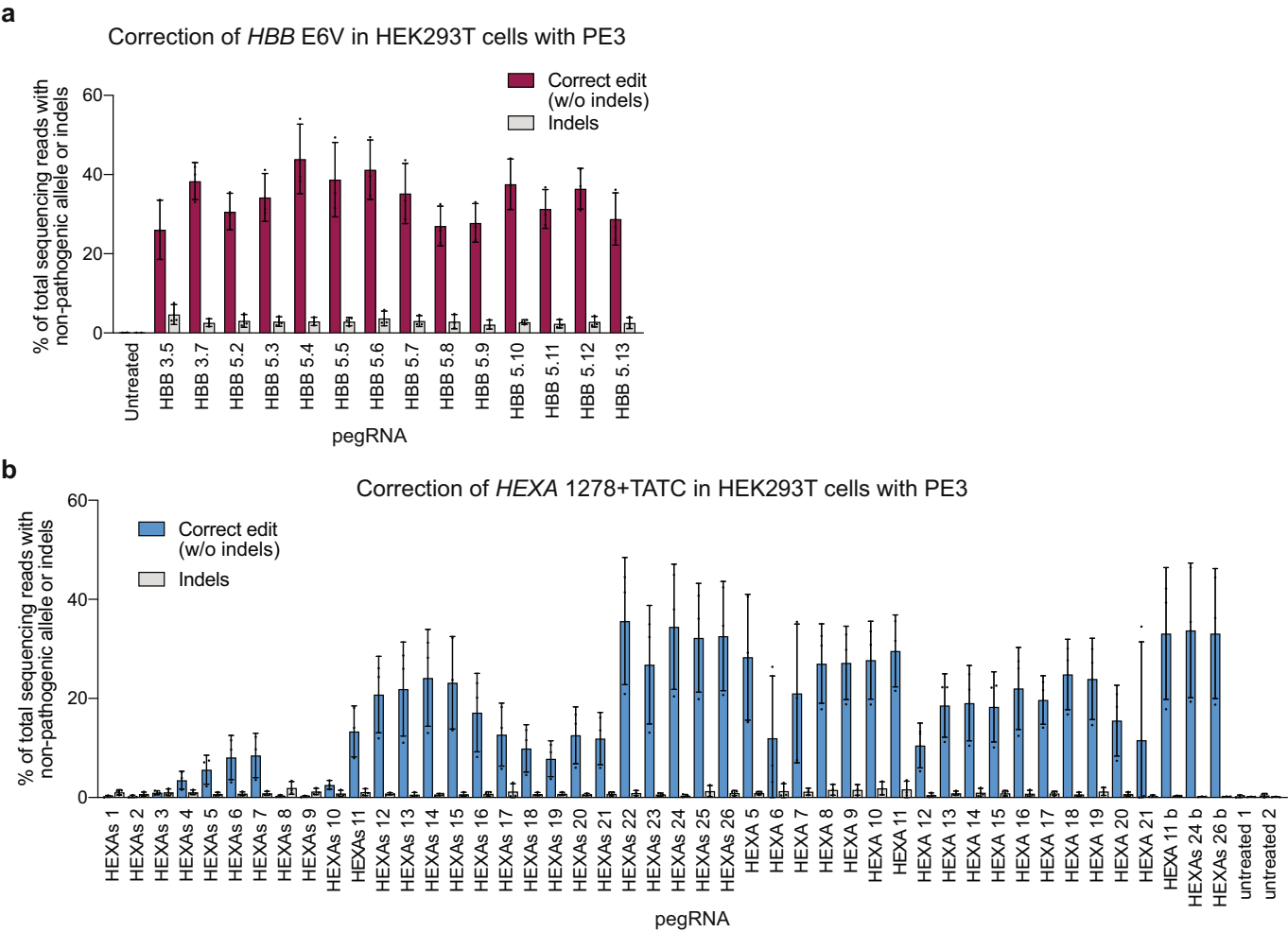


Extended Data Fig. 8 | See next page for caption.

**Extended Data Fig. 8 | Effects of PE2, PE2-dRT, Cas9(H840A) nickase, and dCas9 on cell viability and on transcriptome-wide RNA abundance.**

HEK293T cells were transiently transfected with plasmids encoding PE2, PE2(R110S/K103L), Cas9(H840A) nickase, or dCas9, together with a *HEK3*-targeting pegRNA plasmid. Cell viability was measured for the bulk cellular population every 24 h after transfection for 3 days using the CellTiter-Glo 2.0 assay (Promega). **a**, Viability, as measured by luminescence, at 1, 2, or 3 days after transfection. Mean  $\pm$  s.e.m. of  $n = 3$  independent biological replicates, each performed in technical triplicate. **b**, Percentage editing and indels for PE2, PE2(R110S/K103L), Cas9(H840A) nickase, or dCas9, together with a *HEK3*-targeting pegRNA plasmid that encodes a +5 G-to-A edit. Editing efficiencies were measured on day 3 after transfection from cells treated alongside those used for assaying viability in **a**. Mean  $\pm$  s.d. of  $n = 3$  independent biological replicates. **c–k**, Analysis of cellular RNA, depleted for ribosomal RNA, isolated from HEK293T cells expressing PE2, PE2-dRT, or Cas9(H840A) nickase and a *PRNP*-targeting or *HEXA*-targeting pegRNA. RNAs corresponding to 14,410

genes and 14,368 genes were detected in *PRNP* and *HEXA* samples, respectively. **c–h**, Volcano plot displaying the  $-\log_{10}$  FDR-adjusted  $P$  value versus  $\log_2$ -fold change in transcript abundance for each RNA, comparing PE2 versus PE2-dRT with *PRNP*-targeting pegRNA (**c**), PE2 versus Cas9(H840A) with *PRNP*-targeting pegRNA (**d**), PE2-dRT versus Cas9(H840A) with *PRNP*-targeting pegRNA (**e**), PE2 versus PE2-dRT with *HEXA*-targeting pegRNA (**f**), PE2 versus Cas9(H840A) with *HEXA*-targeting pegRNA (**g**), PE2-dRT versus Cas9(H840A) with *HEXA*-targeting pegRNA (**h**). Red dots indicate genes that show twofold or more changes in relative abundance that are statistically significant (FDR-adjusted  $P < 0.05$ ). **i–k**, Venn diagrams of upregulated and downregulated transcripts (twofold change or more) comparing *PRNP* and *HEXA* samples for PE2 versus PE2-dRT (**i**), PE2 versus Cas9(H840A) (**j**), and PE2-dRT versus Cas9(H840A) (**k**). Values for each RNA-seq condition reflect the mean of  $n = 5$  biological replicates. Differential expression was assessed using a two-sided  $t$ -test with empirical Bayesian variance estimation.



**Extended Data Fig. 9 | PE3-mediated correction of E6V-encoding *HBB* mutation and *HEXA*<sup>1278+TATC</sup> by various pegRNAs. a**, Screen of 14 pegRNAs for correction of the *HBB* E6V-encoding allele in HEK293T cells with PE3. All pegRNAs evaluated convert the mutant *HBB* allele back to wild-type *HBB* without the introduction of any silent PAM mutation. **b**, Screen of 41 pegRNAs for correction of the *HEXA*<sup>1278+TATC</sup> allele in HEK293T cells with PE3 or PE3b.

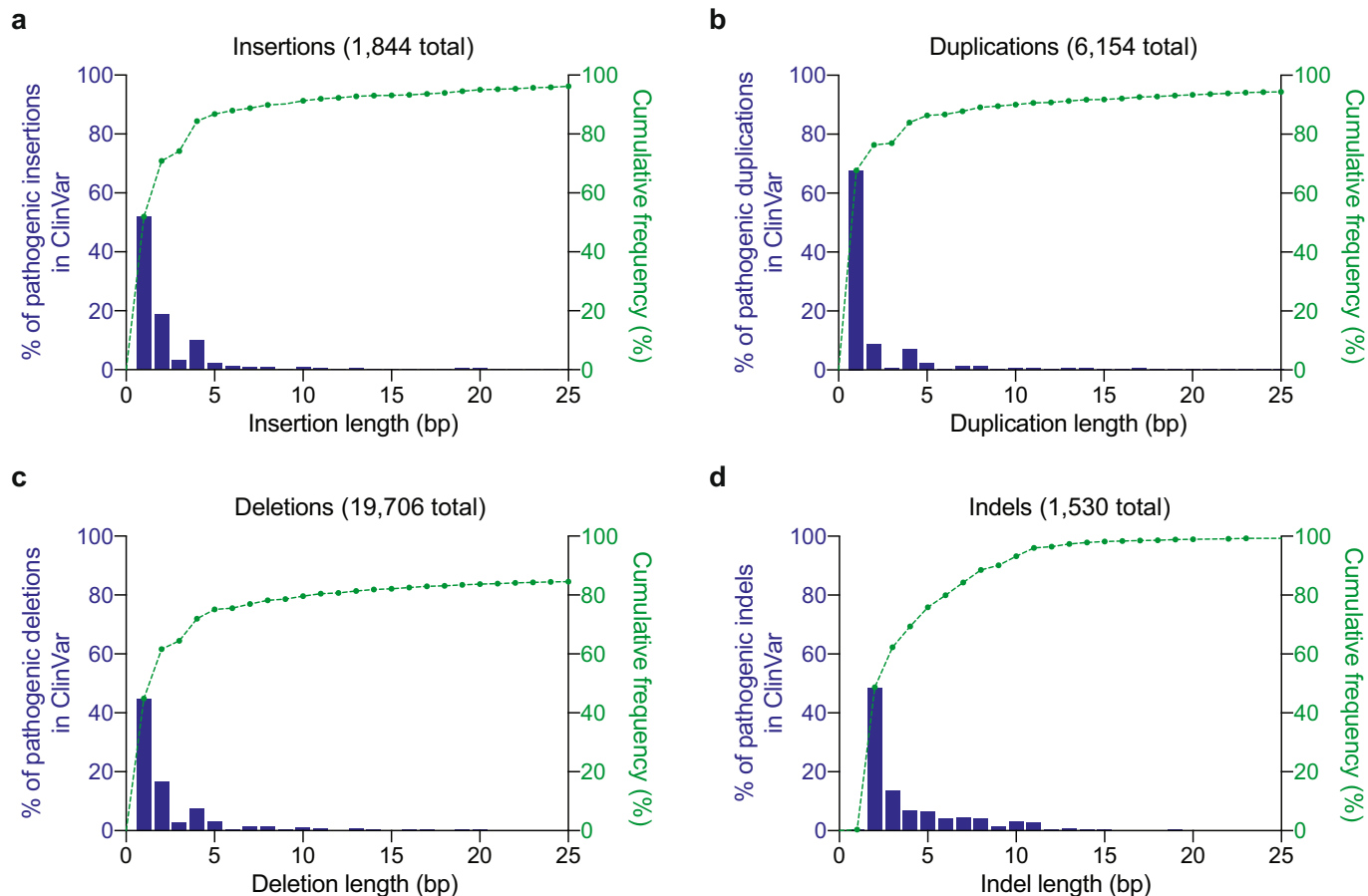
Those pegRNAs labelled HEXAs correct the pathogenic allele by a shifted 4-bp deletion that disrupts the PAM and leaves a silent mutation. Those pegRNAs labelled HEXA correct the pathogenic allele back to wild-type. Entries ending in b use an edit-specific nicking sgRNA in combination with the pegRNA (the PE3b system). Mean  $\pm$  s.d. of  $n = 3$  independent biological replicates.



**Extended Data Fig. 10 | PE3 activity in human cell lines and comparison of PE3 and Cas9-initiated HDR.** **a**, Prime editing in K562 (leukaemic bone marrow), U2OS (osteosarcoma), and HeLa (cervical cancer) cells. **b–e**, Efficiency of generating the correct edit (without indels) and indel frequency for PE3 and Cas9-initiated HDR in HEK293T cells (**b**), K562 cells (**c**), U2OS cells (**d**), and HeLa cells (**e**). Each bracketed editing comparison installs identical edits with PE3 and Cas9-initiated HDR. Non-targeting controls are PE3 and a pegRNA that targets a non-target locus. (**f**) Control experiments with non-targeting pegRNA + PE3, and with dCas9 + sgRNA, compared with wild-type Cas9 HDR experiments confirming that ssDNA donor HDR template, a common contaminant that artificially elevates apparent HDR efficiencies,

does not contribute to the HDR measurements in **a–d**. **g**, Example *HEK3* site allele tables from genomic DNA samples isolated from K562 cells after editing with PE3 or with Cas9-initiated HDR. Alleles were sequenced on an Illumina MiSeq and analysed using CRISPResso2<sup>43</sup>. The reference *HEK3* sequence from this region is at the top. Allele tables are shown for a non-targeting pegRNA negative control, a +1 CTT insertion at *HEK3* using PE3, and a +1 CTT insertion at *HEK3* using Cas9-initiated HDR. Allele frequencies and corresponding Illumina sequencing read counts are shown for each allele. All alleles observed with frequency  $\geq 0.20\%$  are shown. Mean  $\pm$  s.d. of  $n = 3$  independent biological replicates.





**Extended Data Fig. 11 | Distribution by length of pathogenic insertions, duplications, deletions, and indels in the ClinVar database.** The ClinVar variant summary was downloaded from NCBI on 15 July 2019. The lengths of reported insertions, deletions, and duplications were calculated using reference and alternate alleles, variant start and stop positions, or appropriate identifying information in the variant name. Variants that did not report any of the above information were excluded from the analysis. The lengths of

reported indels (single variants that include both insertions and deletions relative to the reference genome) were calculated by determining the number of mismatches or gaps in the best pairwise alignment between the reference and alternate alleles. **a**, Length distribution of insertions. **b**, Length distribution of duplications. **c**, Length distribution of deletions. **d**, Length distribution of indels.

# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☒ ☐ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☒ ☐ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

## Software and code

Policy information about [availability of computer code](#)

Data collection

Illumina Miseq Control software (3.1) was used on the Illumina Miseq sequencers to collect the high-throughput sequencing data

Data analysis

Crispresso2 was used to analyze HTS data for quantifying editing activity at genomic sites. Cell Sorter Software Version 3.0.5 was used for flow cytometry analysis. RNA-seq demultiplexing was performed with bcl2fastq2 version 2.20, and sequences were trimmed with TrimGalore v. 0.6.2. Alignment of RNA-seq reads to the human genome was performed with RSEM version 1.3.1. RNA-seq data output was generated with limma-voom and visualized in R. Frequency, mean, and standard deviations were calculated using GraphPad Prism 8. Custom python scripts provided in Supplementary Note 4 were used to analyze and quantify guide RNA scaffold insertion.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

High-throughput sequencing data have been deposited in the NCBI Sequence Read Archive database under accession code PRJNA565979.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were determined based on literature precedence for genome editing experiments.
Data exclusions	No data was excluded.
Replication	All experiments were repeated at least once. All attempts at replication were successful.
Randomization	Yeast and mammalian cells used in this study were grown under identical conditions; no randomization was used.
Blinding	Yeast and mammalian cells used in this study were grown under identical conditions; blinding was not used.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	HEK293T (ATCC), U2OS (ATCC), K562 (ATCC), HeLa (ATCC).
Authentication	Cells were authenticated by the supplier using STR analysis.
Mycoplasma contamination	All cell lines tested negative for mycoplasma.
Commonly misidentified lines (See <a href="#">ICLAC</a> register)	None used.

## Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	To generate dissociated neuronal cultures, timed-pregnant C57BL/6 mice were provided by Charles River. Pregnant mice were euthanized at E18.5, and tissue for dissociated cultures was harvested from all embryos.
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve samples collected from the field.
Ethics oversight	The Broad IACUC provided ethical guidance.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Flow Cytometry

## Plots

Confirm that:

- ☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☒ All plots are contour plots with outliers or pseudocolor plots.
- ☒ A numerical value for number of cells or percentage (with statistics) is provided.

## Methodology

Sample preparation

2.5 x 10<sup>4</sup> HEK293T cells grown in the absence of antibiotic were seeded on 48-well poly-D-lysine coated plates (Corning). 16–24 h post-seeding, cells were transfected at approximately 70% confluency with 1  $\mu$ L of Lipofectamine 2000 (Thermo Fisher Scientific) according to the manufacturer's protocols and 750 ng of PE2-P2A-GFP plasmid, 250 ng of pegRNA plasmid, and 83 ng of sgRNA plasmid. After 3 days post transfection, cells were washed with phosphate-buffered saline (Gibco) and dissociated using TrypLE Express (Gibco). Cells were then diluted with DMEM plus GlutaMax (Thermo Fisher Scientific) supplemented with 10% (v/v) FBS (Gibco) and passed through a 35- $\mu$ m cell strainer (Corning) prior to sorting. Cells were treated with 3 nM DAPI (BioLegend) 15 minutes prior to sorting.

Instrument

Sony LE-MA900 Cell Sorter

Software

Cell Sorter Software Version 3.0.5 (Sony)

Cell population abundance

Of the surviving single sorted HEK293T cells edited to have HEXA 1278+TATC, 3.02% were homozygous. Of the surviving single sorted HEK293T cells edited to have HBB E6V, 25% were homozygous. Cells were genotyped using next-generation sequencing (Illumina).

Gating strategy

HEK293T cells were initially gated on population using FSC-A/BSC-A (Gate A) and then sorted for singlets using FSC-A/FSC-H (Gate B). Live cells were sorted for by gating for DAPI-negative cells (Gate C). Finally the upper 50% of GFP expressing cells were sorted for using eGFP as the fluorochrome (Gate D).

- ☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

# Chromatin structure dynamics during the mitosis-to-G1 phase transition

<https://doi.org/10.1038/s41586-019-1778-y>

Received: 17 March 2019

Accepted: 2 October 2019

Published online: 27 November 2019

Haoyue Zhang<sup>1</sup>, Daniel J. Emerson<sup>2</sup>, Thomas G. Gilgenast<sup>2</sup>, Katelyn R. Titus<sup>2</sup>, Yemin Lan<sup>3</sup>, Peng Huang<sup>1</sup>, Di Zhang<sup>1,3</sup>, Hongxin Wang<sup>1</sup>, Cheryl A. Keller<sup>4</sup>, Belinda Giardine<sup>4</sup>, Ross C. Hardison<sup>4</sup>, Jennifer E. Phillips-Cremens<sup>2\*</sup> & Gerd A. Blobel<sup>1,3\*</sup>

Features of higher-order chromatin organization—such as A/B compartments, topologically associating domains and chromatin loops—are temporarily disrupted during mitosis<sup>1,2</sup>. Because these structures are thought to influence gene regulation, it is important to understand how they are re-established after mitosis. Here we examine the dynamics of chromosome reorganization by Hi-C after mitosis in highly purified, synchronous mouse erythroid cell populations. We observed rapid establishment of A/B compartments, followed by their gradual intensification and expansion. Contact domains form from the ‘bottom up’—smaller subTADs are formed initially, followed by convergence into multi-domain TAD structures. CTCF is partially retained on mitotic chromosomes and immediately resumes full binding in ana/telophase. By contrast, cohesin is completely evicted from mitotic chromosomes and regains focal binding at a slower rate. The formation of CTCF/cohesin co-anchored structural loops follows the kinetics of cohesin positioning. Stripe-shaped contact patterns—anchored by CTCF—grow in length, which is consistent with a loop-extrusion process after mitosis. Interactions between *cis*-regulatory elements can form rapidly, with rates exceeding those of CTCF/cohesin-anchored contacts. Notably, we identified a group of rapidly emerging transient contacts between *cis*-regulatory elements in ana/telophase that are dissolved upon G1 entry, co-incident with the establishment of inner boundaries or nearby interfering chromatin loops. We also describe the relationship between transcription reactivation and architectural features. Our findings indicate that distinct but mutually influential forces drive post-mitotic chromatin reconfiguration.

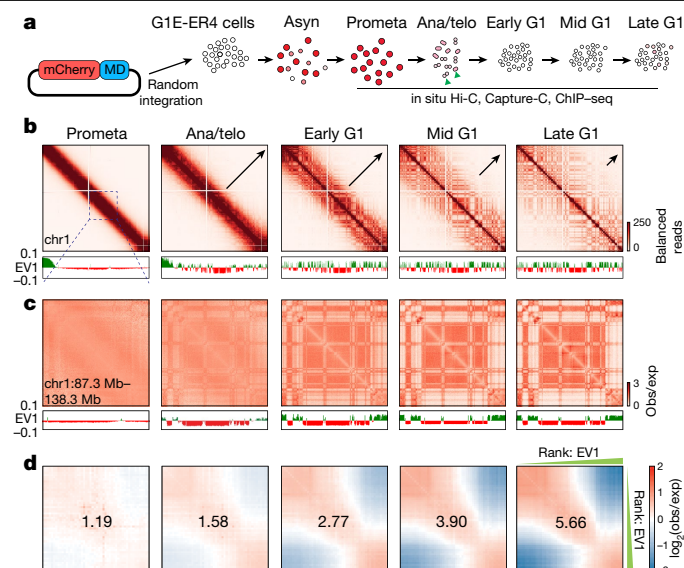
The global restructuring of chromosomal architecture during the progression from mitosis into G1 phase provides an opportunity to examine hierarchies and mechanisms of chromosome organization<sup>3</sup> (Extended Data Fig. 1a). We performed *in situ* Hi-C experiments<sup>4</sup> at defined time points after mitosis following nocodazole-induced prometaphase arrest–release in G1E-ER4 cells, a well-characterized subline<sup>5</sup> of the mouse erythroblast line G1E (Fig. 1a). To ensure maximal purity of cell populations, we used a fluorescence-activated-cell-sorting (FACS)-based isolation strategy based on cell cycle markers and DNA content (Extended Data Fig. 1b, c; Supplementary Methods). *In situ* Hi-C collectively yielded around 2 billion uniquely mapped interactions, with high concordance between biological replicates (Extended Data Fig. 1d–f). Consistent with previous studies, compartments are largely eliminated in prometaphase<sup>1,2</sup> (Fig. 1b). In ana/telophase—the earliest examined interval—compartments are already detectable visually and by eigenvector decomposition, and gain in intensity as cells advance into G1 (Fig. 1b–d; Extended Data Fig. 2a–c). This is consistent with the results of a multiplexed 4C-seq study, which reported the early establishment

of compartments after mitosis<sup>6</sup>. As expected, the A-type compartment is associated with active histone marks<sup>7</sup> (Extended Data Fig. 2d). As cells proceed towards late G1, the characteristic checkerboard pattern of compartments visually expands away from the diagonal, leading to increased interaction frequencies at large (>100 Mb) distance scales (Fig. 1b, Extended Data Fig. 2e, f). Quantification of compartmentalization at different genomic distance scales across all cell cycle stages revealed a progressive gain of compartmentalization among distant (>100 Mb) genomic regions, confirming the expansion of compartments after mitosis (Extended Data Fig. 2g–i; Supplementary Methods). Therefore, a major reconfiguration of genome structure occurs during the prometaphase–G1 phase transition, involving a rapid establishment, progressive strengthening, and expansion of A/B compartments throughout the chromosome.

Next we examined the formation of topologically associating domains (TADs) and nested subTADs after mitosis using 3DNetMod<sup>8</sup>. We identified a total of 8,082 contact domains that are progressively gained from prometaphase to mid G1 (Fig. 2a; Supplementary Table 1).

<sup>1</sup>Division of Hematology, The Children's Hospital of Philadelphia, Philadelphia, PA, USA. <sup>2</sup>Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, USA. <sup>3</sup>Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. <sup>4</sup>Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA, USA. \*e-mail: jcremins@seas.upenn.edu; blobel@email.chop.edu

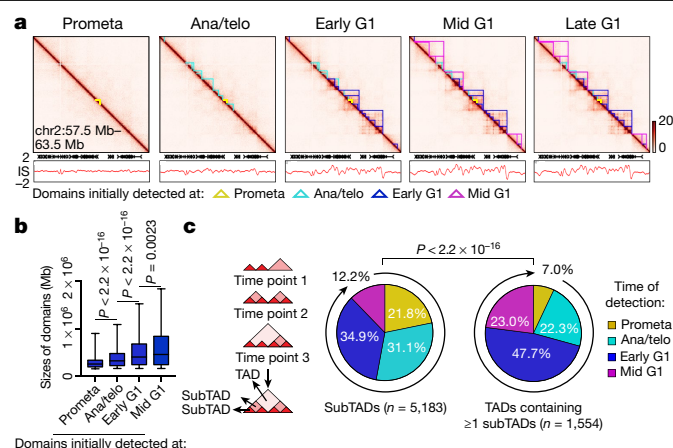




**Fig. 1 | Early appearance and progressive strengthening and expansion of A/B compartments after mitosis. a**, Schematic showing the reporter gene encoding mCherry fused to the mouse cyclin B mitotic degradation domain (mCherry-MD) and the expected mCherry signal at each cell cycle stage. Green arrowheads indicate sorting of cells in anaphase or telophase (ana/telo). Asyn, asynchronous; prometa, prometaphase. **b**, Hi-C contact maps showing the restoration of chromatin A/B compartments of chromosome 1 (chr1) after mitosis, along with genome browser tracks showing eigenvector 1 values. Bin size, 250 kb. Arrows indicate expansion of compartments. **c**, A magnified view (chr1:87.3 Mb–138.3 Mb) of **b** revealing the clear plaid-like compartment pattern in ana/telophase. **d**, Saddle plots showing the genome-wide compartment strength over time.

Establishment of boundaries and enrichment of intradomain interactions were observed at newly emerging domains, thereby validating our domain-calling approach (Extended Data Fig. 3a–e). Previous studies have reported a complete loss of domains in prometaphase<sup>1,2</sup>. However, despite considerable attenuation, residual domain- and boundary-like structures are still detectable visually and algorithmically in prometaphase cells (Extended Data Fig. 3f). To rule out contamination by G1 cells as a cause of prometaphase domain detection, we simulated *in silico* admixing with up to 20% of G1 chromosomes. Even a G1 contribution of 20%—which far exceeds the observed interphase cell contamination of up to 2%—did not reproduce patterns observed in prometaphase (Extended Data Fig. 3f–h); this suggests that prometaphase domain- and boundary-like features are not due to the presence of G1 phase cells. Residual domain boundaries in prometaphase are enriched with active histone marks and transcription start sites<sup>9</sup> (Extended Data Fig. 3i, j).

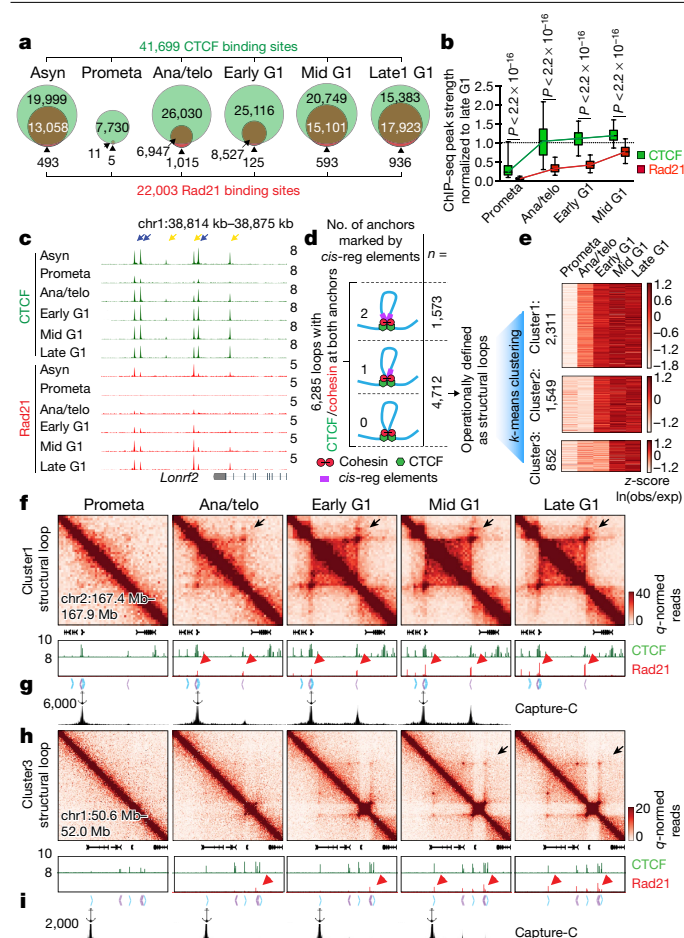
Formation of nested domain structures may occur through the convergence of previously emerged subTADs (bottom-up), the partitioning of initially formed TADs into subTADs (top-down), or the simultaneous appearance of both contact domain types (Extended Data Fig. 4a). On average, contact domains that are established at time points later in G1 are larger than those called at earlier stages of the cell cycle (Fig. 2a, b); this observation favours the bottom-up formation scenario. To further test this model, we categorized all contact domains into 2,899 TADs and 5,183 subTADs on the basis of their hierarchical organization (Fig. 2c). Notably, higher proportions of subTADs are detected in prometaphase or ana/telophase compared to the TADs that encompass them, which suggests that subTADs tend to assemble more rapidly (Fig. 2c). Once established, the majority of TADs remain unchanged without further subdivisions, disfavours the ‘top-down’ model (Extended Data Fig. 4b). By contrast, 85.4% and 69.1% of subTADs called in prometaphase and ana/telophase, respectively, converge into larger domains



**Fig. 2 | Contact domains develop from the bottom up after mitosis. a**, Hi-C contact maps coupled with insulation score tracks (chr2: 57.5 Mb–63.5 Mb). Domains emerging at each stage of the cell cycle are demarcated by colour-coded lines. Bin size, 10 kb. Colour bar denotes  $q$ -normed reads. **b**, Sizes of domains newly detected at prometaphase ( $n = 1,528$ ), ana/telophase ( $n = 2,394$ ), early G1 ( $n = 2,995$ ) and mid G1 ( $n = 1,165$ ). For all box plots, centre lines denote medians; box limits denote 25th–75th percentile; whiskers denote 5th–95th percentile.  $P$  values were calculated by a two-sided Mann–Whitney  $U$ -test. **c**, Left, schematic showing the partition of domains into TADs or subTADs. TADs are domains that are not encompassed by any other domains; subTADs are domains that are completely encompassed by other domains. Right, pie charts of the cell cycle distribution of subTADs and TADs that contain at least one subTAD based on their time of emergence.  $P$  values were calculated using a two-sided Fisher’s exact test (prometaphase + ana/telophase compared with early G1 + mid G1).

during later stages (Extended Data Fig. 4c). In line with subTAD merging, we observed gains in contacts across subTAD boundaries over time (Extended Data Fig. 4d). Accordingly, a substantial portion of subTAD boundaries detected at prometaphase exhibit increased insulation scores (indicative of reduced insulation), whereas for most TAD boundaries, insulation scores decrease as cells progressed from prometaphase into G1 (Extended Data Fig. 4e). Independent algorithms yielded similar trends of subTAD merging after mitosis<sup>8,10</sup> (Extended Data Fig. 4f–m). Together, these analyses suggest a ‘bottom-up’ model of hierarchical domain reorganization during the prometa-to-G1 phase transition.

A loop-extrusion model has been proposed to explain the formation of TADs and chromatin loops, wherein the cohesin complex extrudes the chromatid until it encounters pairs of convergently oriented CTCF-binding sites<sup>11,12</sup>. Because cell cycle dynamics of loop formation as well as CTCF and cohesin binding could inform this (or alternative) models, we surveyed the chromatin-binding profiles of CTCF and cohesin by chromatin-immunoprecipitation followed by sequencing (ChIP-seq). We generated highly concordant replicates (Extended Data Fig. 1g, h) and identified 41,699 CTCF-binding sites and 22,003 binding sites for Rad21, a cohesin subunit (Supplementary Table 2). Approximately 88.7% (19,520) of Rad21 peaks were co-occupied by CTCF. Notably, around 18.6% (7,741) of CTCF peaks are reproducibly detected in prometaphase cells, indicating that a considerable amount of CTCF remains bound to mitotic chromatin (Extended Data Fig. 5a, c, d). Previous reports have described varying degrees of CTCF mitotic retention<sup>13,14</sup>. Unlike CTCF, Rad21 failed to show localized chromatin binding during prometaphase (Extended Data Fig. 5b–d). Motif scan and genomic distribution analysis failed to identify distinct features associated with CTCF peaks present in both interphase and mitosis (IM-peaks) (Extended Data Fig. 5e, f). However, IM-peaks tend to be more tissue invariant and are more likely to be co-occupied by Rad21 during interphase (Extended Data Fig. 5f). CTCF and cohesin resume chromatin occupancy after mitosis with markedly different kinetics.



**Fig. 3 | Focal accumulation of cohesin is delayed compared to that of CTCF and coincides with structural loop formation.** **a**, Venn diagrams showing the distribution of CTCF and Rad21 ChIP-seq peaks across cell cycle stages. **b**, Box plots showing the recovery rate of CTCF ( $n = 33,306$ ) and Rad21 ( $n = 18,859$ ) peaks. Peaks absent from late G1 were omitted from the analysis. For all box plots, centre lines denote medians; box limits denote 25th–75th percentile; whiskers denote 5th–95th percentile. *P* values were calculated using a two-sided Mann–Whitney *U*-test. **c**, Genome browser tracks of CTCF and Rad21 at the *LONRF2* locus across cell cycle stages.  $n = 2$ –3 biological replicates. Blue and yellow arrows indicate IM- and interphase-only (IO)-CTCF binding sites, respectively. **d**, Schematic depicting the classification of loops. All loops with CTCF/cohesin co-occupancy at both anchors were subdivided into those with 0, 1 or 2 anchors marked by *cis*-regulatory elements. Those with 0 or 1 were operationally defined as structural loops. **e**, Heat map showing the result of *k*-means clustering on the 4,712 structural loops. **f**, Hi-C contact maps showing a representative region that contains a cluster 1 structural loop (chr2: 167.4 Mb–167.9 Mb, black arrows), along with genome browser tracks of CTCF and Rad21 ChIP-seq profiles. Rad21 peaks at two loop anchors are indicated by red arrowheads. Chevron arrows highlight positions and orientations of CTCF sites at the loop anchors. Bin size, 10 kb. **g**, Capture-C interaction profile of the same region as shown in **f**.  $n = 3$  biological replicates. The anchor symbol shows position of the capture probe. **h, i**, similar to **f, g**, showing a representative region that contains a cluster 3 (slowly emerging) structural loop (chr1: 50.6 Mb–52.0 Mb, black arrows).

The majority of CTCF peaks are immediately restored in ana/telophase, whereas Rad21 peaks appear much more gradually (Fig. 3a–c; Extended Data Fig. 5g–i). Delayed nuclear import as well as chromatin loading and/or movement along the chromatid could account for the slow focal accumulation of cohesin after mitosis. We performed live-cell imaging on asynchronous G1E-ER4 cells that endogenously express mCherry-tagged CTCF or mCherry-tagged SMC3 (a cohesin subunit)

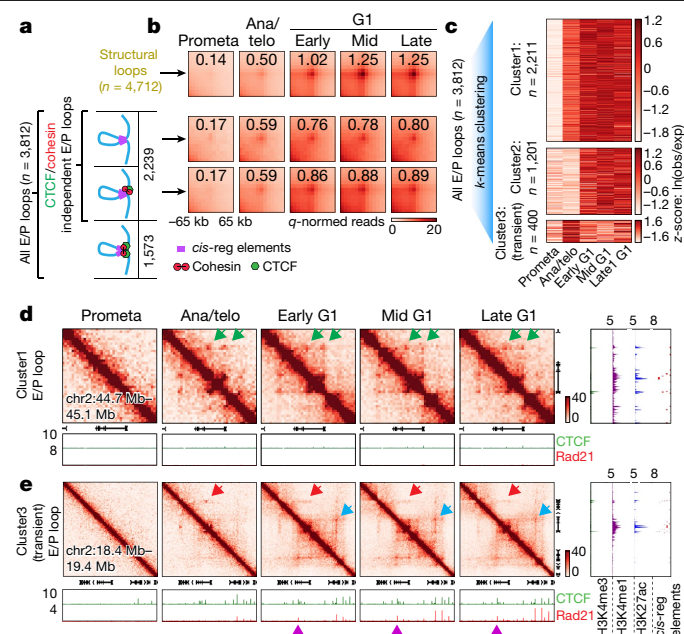
(Extended Data Fig. 5j). Consistent with the ChIP-seq data and a previous report<sup>15</sup>, CTCF rapidly accumulates on telophase chromosomes, whereas SMC3 is excluded from chromosomes during metaphase, telophase and cytokinesis (Extended Data Fig. 5k). Moreover, after G1 entry, nuclear import of SMC3 is also slower compared to that of CTCF (Extended Data Fig. 5k, l). These results suggest that the delayed kinetics of focal cohesin accumulation might be a composite of nuclear import, association with chromatin, and migration along the chromatid.

The transient decoupling of cohesin from CTCF during mitotic exit offers the opportunity to separately assess their roles in post-mitotic loop formation. Using a modified HICCUPS algorithm we identified 13,317 chromatin loops, progressively gained from prometaphase to late G1, with highly concordant loop strength between biological replicates (Extended Data Fig. 6a–c; Supplementary Table 3). Of these loops, 6,285 (about 47.2%) contain CTCF and cohesin co-occupied sites at both anchors (Fig. 3d). These loops were further filtered to eliminate interactions between putative *cis*-regulatory elements (for example, enhancer–promoter loops), resulting in 4,712 operationally defined ‘structural’ loops (Fig. 3d). To investigate how fast structural loops are formed we performed *k*-means clustering, which revealed three clusters with distinct formation dynamics (Fig. 3e). Cluster 1 loops display strong interactions in ana/telophase, whereas the formation of cluster 2 and 3 loops is delayed (Fig. 3e, f, h; Extended Data Fig. 6d, e). Analysis by Capture-C validated the differential dynamics of structural loops at two representative loci (Fig. 3g, i). Notably, anchors of cluster 1 loops show enrichment of Rad21 at ana/telophase, whereas anchors of cluster 2 and 3 loops acquire Rad21 more gradually (Fig. 3f, h; Extended Data Fig. 6d, e). By contrast, CTCF is rapidly enriched at anchors of all three loop clusters (Fig. 3f, h; Extended Data Fig. 6d, e). The strengths of structural loops are highly correlated with ChIP-seq signals of Rad21 at their anchors over time, but significantly less so with those of CTCF (Extended Data Fig. 6f). Late-occurring structural loops are significantly larger than earlier ones, suggesting a correlation between size and time to formation (Extended Data Fig. 6g). Together, our results reveal three clusters of structural loops with distinct formation dynamics, and suggest that the accumulation of cohesin—but not CTCF—is rate-limiting for the formation of structural loops after mitosis.

Stripes in the contact maps are thought to reflect interactions between a single locus and a continuum of genomic regions, and are considered as evidence for the loop extrusion model<sup>17</sup>. Using a modified statistical modelling approach<sup>17</sup>, we identified 1,775 stripes genome-wide. The majority of them contain inwardly oriented CTCF sites at their anchors (Extended Data Fig. 7a). Notably, these striped contacts grow directionally over time but display punctuated enrichment at select CTCF sites (Extended Data Fig. 7b, d). This is consistent with an extrusion mechanism in which some CTCF-binding sites serve as obstacles to cohesin processivity. We also observed blockage of stripe extension that correlates with the presence of strong CTCF-binding sites, resulting in the formation of structural loops at the far end of the stripes (Extended Data Fig. 7b). Together, our data are consistent with dynamic loop extrusion after mitosis. Stripe-like patterns that appear rapidly with little or no further growth were also observed, and are discussed below (Extended Data Fig. 7c, e, f).

Next we investigated interactions between *cis*-regulatory elements. We identified 3,812 chromatin loops with both anchors marked by promoters or putative enhancers, which we termed E/P loops (Fig. 4a). This number is probably an underestimate because short range E/P loops can escape detection. Notably, a considerable portion (approximately 58.7%, 2,239) of E/P loops have only one or no anchor that co-localizes with CTCF and cohesin co-occupied sites, suggesting that E/P loops may form by a mechanism other than CTCF/cohesin-mediated loop extrusion (Fig. 4a). These seemingly CTCF/cohesin independent E/P loops are intensified significantly faster than structural loops (Fig. 4b, Extended Data Fig. 6h). Note that the faster formation of E/P loops compared to structural loops is not explained by differences in loop





**Fig. 4 | *cis*-Regulatory contacts are established rapidly after mitosis and can be transient.** **a**, Schematic depicting the classification of loops. E/P loops were subdivided into those with 0, 1 or 2 anchors containing CTCF/cohesin co-occupied sites. Those with 0 or 1 anchor co-occupied by CTCF/cohesin were classified as E/P loops independent from CTCF and cohesin. **b**, Aggregated peak analysis of CTCF/cohesin independent E/P loops (middle and bottom) in comparison to structural loops (top). Bin size, 10 kb. Numbers indicate average loop strength:  $\ln(\text{observed}/\text{expected})$ . **c**, Heat map of *k*-means clustered E/P loops. **d**, Hi-C contact maps of a representative region (chr2:44.7 Mb–45.1 Mb) containing cluster 1 E/P loops (green arrows), coupled with browser tracks of CTCF and Rad21 occupancy. Bin size, 10 kb. The colour bar denotes *q*-normed reads. Tracks of H3K4me3, H3K4me1, H3K27ac and annotations of *cis*-regulatory elements were from asynchronously growing GIE-ER4 cells. **e**, Similar to **d**, a representative region (*Commd3* locus, chr2:18.4 Mb–19.4 Mb) containing a cluster 3 (transient) E/P loop (red arrows). Blue arrows denote the formation of a downstream, potentially interfering structural loop. Purple arrowheads indicate CTCF/cohesin binding at the potentially interfering structural loop anchor.

size (Extended Data Fig. 6i). Accordingly, among loops established in ana/telophase, about 69.3% are E/P loops whereas only 11.6% are structural loops (Extended Data Fig. 6j). These trends are reversed in mid G1 (18.4% E/P and 42.3% structural loops, respectively). Hence, E/P loops may not require CTCF and cohesin, and can be rebuilt faster than structural loops after mitosis.

Clustering all E/P loops on the basis of their time of enrichment yielded at least three classes with distinct post-mitotic formation kinetics. Cluster 1 (2,211, 58%) E/P contacts are rapidly enriched in ana/telophase, whereas cluster 2 contacts (1,201, 31.5%) form in early G1 (Fig. 4c, d; Extended Data Fig. 8a, b). We also discovered a third cluster (400, 10.5%) of E/P loops that peak early in ana/telophase and gradually diminish in G1 (Fig. 4c, e; Extended Data Fig. 8c, d, f). We independently validated this transient nature between certain *cis*-regulatory elements by Capture-C at the two manually identified loci *Pde12* and *Morc3* (Extended Data Fig. 8c, e). In an effort to understand the mechanisms that underlie this subset of transient E/P loops, we noticed that approximately 55% of them span either a boundary or an anchor of a nearby structural loop that is established later in G1 (Fig. 4e, Extended Data Fig. 8c). Moreover, these boundaries and loop anchors within cluster 3 E/P loops display more substantial insulation compared to those within clusters 1 or 2 (Extended Data Fig. 8g). We therefore speculate that emerging boundaries or nearby structural loops may interfere with E/P loops (Extended Data Fig. 1a). To test this hypothesis, we set out to assay cluster 3 E/P loop dynamics after perturbing the nearby structural

loop. We focused on the interaction between the *Commd3* promoter and a distal *cis*-regulatory element. We deleted the CTCF core motif of a potential interfering structural loop anchor, resulting in the abrogation of CTCF and Rad21 binding (Extended Data Fig. 8f, h, i). Notably, in the mutant cells, interactions between the *Commd3* promoter and the distal *cis*-regulatory element are prolonged after mitosis, compared to controls (Extended Data Fig. 8j–l). These results provide a precedent for a dynamic interplay between structural and E/P loops. However, insulation between regulatory elements is unlikely to fully explain the transient nature of cluster 3 E/P loops, because only around 55% of them span boundaries or interfering loop anchors. Additional mechanisms, such as competition between regulatory elements, may also contribute to the transient nature of cluster 3 E/P loops. In summary, we identified a special class of transient E/P contacts after mitosis, which may in some cases be broken by CTCF and cohesin.

To explore the relationship between chromatin organization and transcription activation<sup>18</sup> after mitosis, we carried out Pol II ChIP-seq<sup>19</sup> (Extended Data Fig. 1i). Transcription is largely silenced in prometaphase, but rapidly reinitiates in ana/telophase and positively correlates with A-type compartments (Extended Data Fig. 9a, b). Collectively, we identified 7,535 active genes after mitosis (Supplementary Table 4). Genes display comparable reactivation dynamics regardless of whether they are located in domains called at early or later stages of the cell cycle, suggesting that domain formation may exert only limited influence on gene reactivation after mitosis (Extended Data Fig. 9c). We then stratified active genes on the basis of their Pol II occupancy over time through principal component analysis<sup>19</sup>. In a previous study we observed that a large fraction of genes acquires strong Pol II occupancy early after mitosis, followed by a reduction in signal intensity. This ‘spike’ in gene reactivation manifests as the first principal component (PC1) and separates ‘spiking’ genes from late-activating genes<sup>19</sup>. Similarly, the current data recapitulate this transient hyperactivation as represented by PC1 (Extended Data Fig. 9d–f). To examine the relationship between gene spiking and E/P loop formation, we began by stratifying all active genes on the basis of whether or not they are positioned at E/P loop anchors (Extended Data Fig. 9g, h). In general, formation of E/P loops is positively correlated with Pol II occupancy over time (median Pearson  $r \approx 0.65$ ). Additionally, we found that genes at cluster 3 E/P loops are more likely to display post-mitotic transcriptional spiking compared to those at cluster 1 or 2 loops, or those with no detectable E/P loops (Extended Data Fig. 9i, j). Regarding genes associated with cluster 1 or 2 E/P loops, activation was also positively correlated with loop strength over time (median Pearson  $r \approx 0.67$ ). These results suggest that transient E/P loops may contribute to post-mitotic gene spiking. However, a caveat to this interpretation is that a much larger number of genes spike than are associated with transient E/P loops. This suggests that E/P contacts cannot be solely responsible for spiking in post-mitotic transcription. Nonetheless, although the causal relationship between gene spiking and transient E/P loops remains uncertain, the overall positive correlation between E/P loop strength and Pol II occupancy over time suggests a potential role of E/P contacts in transcription after mitosis.

We exploited the natural transition from a relatively unorganized state (prometaphase) into fully established chromatin organization late in G1 to interrogate mechanisms by which chromatin is hierarchically organized (Extended Data Fig. 1a). We showed that A/B compartmentalization was disrupted in prometaphase despite histone marks being largely maintained<sup>20</sup>. We also show that local (around 10 Mb) compartmentalization of chromatin initiates rapidly after mitosis, and continues to expand and increase in strength. Study of the cell cycle dynamics of chromatin also enabled the testing of predictions made by the loop extrusion model. First, small TADs and structural loops are formed more quickly than larger ones. Second, stripes in the contact maps increase in length over time. Third, based on the kinetics of CTCF and cohesin deposition on chromatin, it is clear that CTCF does not form detectable loops without cohesin even though it can form

multimers<sup>21</sup>. However, it is possible that CTCF pairs with itself—or with other factors such as YY1<sup>22,23</sup>—to facilitate the establishment of contacts among *cis*-regulatory elements, such as those observed at early time points independently of cohesin.

Our integrative analysis of loops and histone-modification profiles reveals a group of E/P loops that can be independent from CTCF and cohesin co-binding. A distinctive feature of E/P loops is their fast appearance compared to structural loops. It is possible that E/P contacts form via collisions of chromatin regions with similar epigenetic states. This is supported by our observation that their post-mitotic recovery rate positively correlates with the intensity of active histone marks at anchors (Extended Data Fig. 8m). It is noteworthy that 16.4% of stripe-like structures that lack inwardly oriented CTCF display only little or no further growth during G1 phase and are highly enriched for histone H3 Lys27 acetylation at their anchors (Extended Data Fig. 7c, e, f). Loop extrusion is unlikely to account for these types of stripe-shaped contact. Instead, these contacts might represent small compartments, defined by local enrichment of transcription factors and chromatin modifications<sup>24</sup>. Similarly, transient E/P loops might result from less discriminatory affinity among regions with similar chromatin states. In summary, our findings describe a dynamic hierarchical framework of post-mitotic chromatin configuration that supports a bottom-up model for the formation of contact domains, implicates CTCF and cohesin in post-mitotic loop extrusion, and identifies extrusion independent pathways that lead to compartmentalization and contacts of *cis*-regulatory networks.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1778-y>.

1. Naumova, N. et al. Organization of the mitotic chromosome. *Science* **342**, 948–953 (2013).
2. Gibcus, J. H. et al. A pathway for mitotic chromosome formation. *Science* **359**, eaao6135 (2018).
3. Nagano, T. et al. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* **547**, 61–67 (2017).

4. Rao, S. S. P. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
5. Weiss, M. J., Yu, C. & Orkin, S. H. Erythroid-cell-specific properties of transcription factor GATA-1 revealed by phenotypic rescue of a gene-targeted cell line. *Mol. Cell. Biol.* **17**, 1642–1651 (1997).
6. Dileep, V. et al. Topologically associating domains and their long-range contacts are established during early G1 coincident with the establishment of the replication-timing program. *Genome Res.* **25**, 1104–1113 (2015).
7. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
8. Norton, H. K. et al. Detecting hierarchical genome folding with network modularity. *Nat. Methods* **15**, 119–122 (2018).
9. Li, B. et al. A comprehensive mouse transcriptomic bodymap across 17 tissues by RNA-seq. *Sci. Rep.* **7**, 4200 (2017).
10. Yu, W., He, B. & Tan, K. Identifying topologically associating domains and subdomains by Gaussian Mixture model And Proportion test. *Nat. Commun.* **8**, 535 (2017).
11. Sanborn, A. L. et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl Acad. Sci. USA* **112**, E6456–E6465 (2015).
12. Fudenberg, G. et al. Formation of chromosomal domains by loop extrusion. *Cell Reports* **15**, 2038–2049 (2016).
13. Oomen, M. E., Hansen, A. S., Liu, Y., Darzacq, X. & Dekker, J. CTCF sites display cell cycle-dependent dynamics in factor binding and nucleosome positioning. *Genome Res.* **29**, 236–249 (2019).
14. Owens, N. et al. CTCF confers local nucleosome resiliency after DNA replication and during mitosis. *eLife* **8**, e47898 (2019).
15. Cai, Y. et al. Experimental and computational framework for a dynamic protein atlas of human cell division. *Nature* **561**, 411–415 (2018).
16. Hughes, J. R. et al. Analysis of hundreds of *cis*-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.* **46**, 205–212 (2014).
17. Vian, L. et al. The energetics and physiological impact of cohesin extrusion. *Cell* **173**, 1165–1178.e20 (2018).
18. Rowley, M. J. et al. Evolutionarily conserved principles predict 3D chromatin organization. *Mol. Cell* **67**, 837–852.e7 (2017).
19. Hsiung, C. C.-S. et al. A hyperactive transcriptional state marks genome reactivation at the mitosis-G1 transition. *Genes Dev.* **30**, 1423–1439 (2016).
20. Behera, V. et al. Interrogating histone acetylation and BRD4 as mitotic bookmarks of transcription. *Cell Rep.* **27**, 400–415.e5 (2019).
21. Yusufzai, T. M., Tagami, H., Nakatani, Y. & Felsenfeld, G. CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Mol. Cell* **13**, 291–298 (2004).
22. Weintraub, A. S. et al. YY1 is a structural regulator of enhancer–promoter loops. *Cell* **171**, 1573–1588 (2017).
23. Beagan, J. A. et al. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. *Genome Res.* **27**, 1139–1152 (2017).
24. Schwarzer, W. et al. Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**, 51–56 (2017).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

All figures include publicly available data. The Hi-C, Capture-C and ChIP-seq data generated and analysed in this study are deposited in the Gene Expression Omnibus repository under accession number GSE129997 for public access. Additional external ChIP-seq data previously reported are available at: H3K27ac (GSE61349)<sup>25</sup>, H3K4me1 (GSM946535)<sup>26</sup>, H3K4me3 (GSM946533)<sup>26</sup>, H3K36me3 (GSM946529)<sup>26</sup> and H3K9me3 (GSM946542)<sup>26</sup>. CTCF peak files from 13 different tissues are available through the ENCODE project (<https://www.encode-project.org/>) with accession numbers ENCFF001LFU, ENCFF001LHE, ENCFF001LHY, ENCFF001LJL, ENCFF001LKO, ENCFF001LMN, ENCFF001LNK, ENCFF001LOR, ENCFF001LPI, ENCFF001LQB, ENCFF001LQS, ENCFF001LSE and ENCFF001LSW.

## Code availability

Code used in this study is available upon request from the corresponding authors.

25. Dogan, N. et al. Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenetics Chromatin* **8**, 16 (2015).

26. Wu, W. et al. Dynamic shifts in occupancy by TAL1 are guided by GATA factors and drive large-scale reprogramming of gene expression during hematopoiesis. *Genome Res.* **24**, 1945–1962 (2014).

**Acknowledgements** We thank members of the Blobel and Phillips-Cremins laboratories for discussions, E. Apostolou and J. Dekker for discussing data before publication, L. Mirny for insights, the CHOP flow core facility staff and A. Stout for expert technical support. This work was supported by grants R37DK058044 to G.A.B., R24DK106766 to G.A.B. and R.C.H., U01HL129998A to J.E.P.-C. and G.A.B., The New York Stem Cell Foundation to J.E.P.-C., the NIH Director's New Innovator Award from the National Institute of Mental Health (1DP2MH11024701; J.E.P.-C.), and a generous gift from the DiGaetano family to G.A.B. J.E.P.-C. is a New York Stem Cell Foundation (NYSCF) Robertson Investigator. We acknowledge support from the Spatial and Functional Genomics program at The Children's Hospital of Philadelphia.

**Author contributions** H.Z., J.E.P.-C. and G.A.B. conceived the study and designed experiments. H.Z. performed experiments with help from P.H., H.W., C.A.K., B.G. and R.C.H. D.J.E. performed initial Hi-C data pre-processing and domain calling. H.Z. performed A/B compartment and ChIP-seq related analysis with help from Y.L. T.G.G. performed loop calling, classification and clustering, and aggregated peak analysis and aggregated domain analysis. K.R.T. performed stripe calling related analysis. D.Z. contributed to Capture-C related analysis. H.Z., J.E.P.-C. and G.A.B. wrote the paper with input from all authors.

**Competing interests** The authors declare no competing interests.

## Additional information

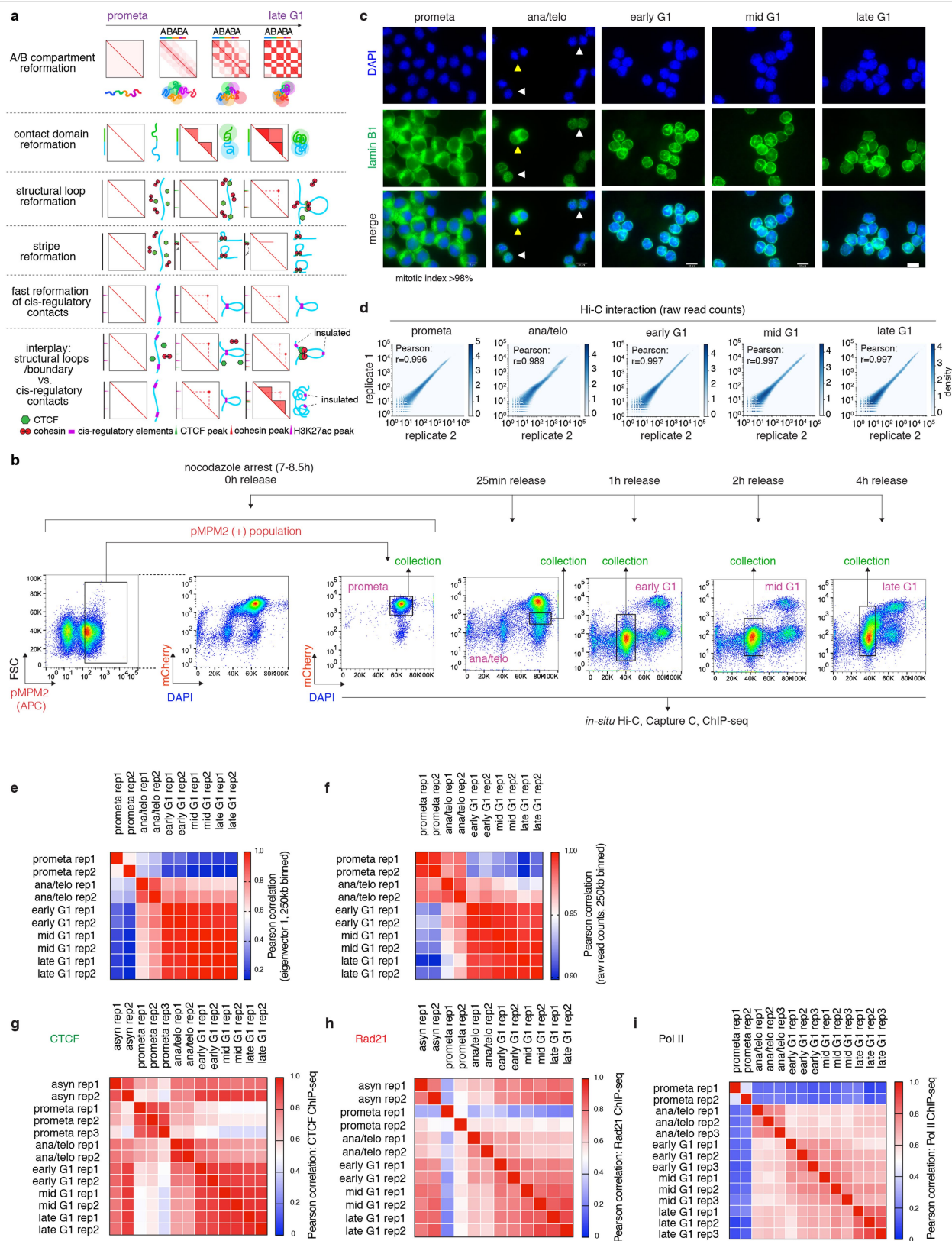
**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1778-y>.

**Correspondence and requests for materials** should be addressed to J.E.P.-C. and G.A.B.

**Peer review information** *Nature* thanks David Gilbert and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.





**Extended Data Fig. 1** | See next page for caption.

**Extended Data Fig. 1 | Models, experimental workflow and data quality control.**

**a**, From top to bottom: schematic illustration of the early emergence, gradual intensification and expansion of A/B compartments (checkerboards) from prometaphase to late G1 phase, coupled with schematics of chromatin organization; subTADs (small triangles) emerge first after mitotic exit, followed by convergence into a TAD (big triangle); formation of a structural loop coincides with the positioning of cohesin, but not CTCF after mitosis; the gradual extrusion of cohesin complex along DNA fibre from one anchor point with CTCF, reflected as enrichment of interactions between the anchor and a continuum of DNA loci on the contact map; fast formation of E/P loops after mitosis; the interplay between transient E/P loops and boundaries or structural loops. **b**, The experimental workflow. Representative flow cytometry plots showing the nocodazole arrest–release strategy based on pMPM2 (prometaphase), mCherry–MD signal, and DNA content (DAPI) staining. Similar observations were made in more than 5 independent experiments.

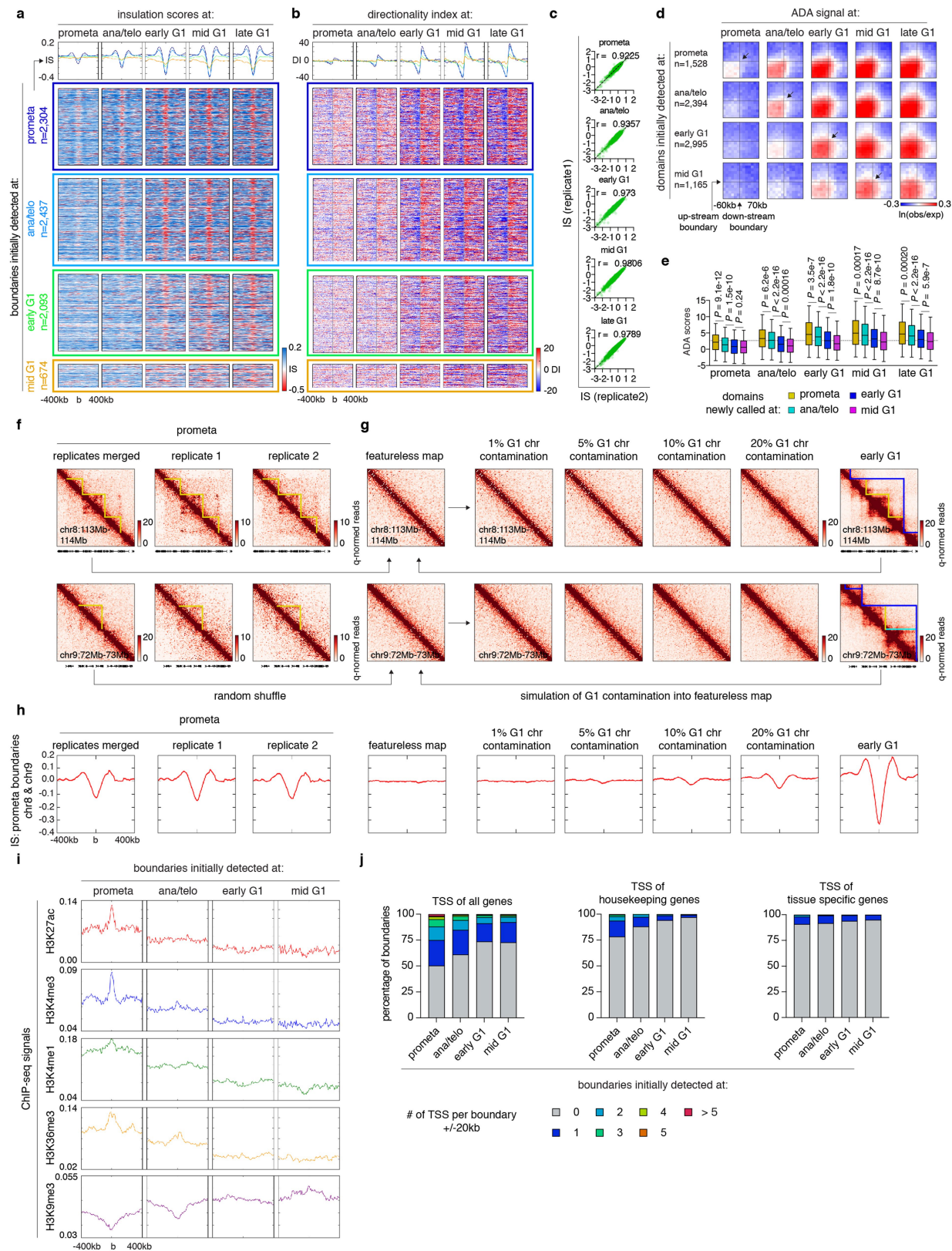
**c**, Representative images showing DAPI and lamin B1 staining of FACS-purified cells across all stages of the cell cycle. Similar observations were made in 2 independent experiments. The mitotic index of prometaphase cells after FACS purification is on average greater than 98%. Yellow and white arrowheads indicate anaphase and telophase cells, respectively. Scale bar, 10  $\mu$ m. **d**, Hexbin plots showing the high correlation of Hi-C raw read counts between two biological replicates across all stages of the cell cycle. Bin size, 250 kb. **e**, Heat map showing the Pearson correlation among all Hi-C samples, based on the eigenvector 1 of 250 kb bins. **f**, Heat map showing the Pearson correlation among all Hi-C samples based on raw read counts. Bin size, 250 kb. **g–i**, Heat maps showing Pearson correlation of CTCF (**g**), Rad21 (**h**) and Pol II (**i**) ChIP-seq data among all samples. Note the overall high replicate concordance. Low correlation coefficients among replicates were only observed in samples with low signal-to-noise ratios—for example, in prometaphase.



**Extended Data Fig. 2 | Compartment strengthening and expansion from ana/telophase throughout late G1.** **a**, Saddle plots showing the progressive gain of compartment strength over time in two biological replicates. **b**, Schematic showing the calculation of compartment strength. **c**, Line graphs showing the progressive increase of compartment strength of each individual chromosome (represented by dots) in two biological replicates. **d**, Heat map showing the genome-wide Spearman correlation coefficients between eigenvector 1 values and asynchronous-cell-derived ChIP-seq signals for the indicated histone marks. **e**, Plots of chromosome-averaged distance-dependent contact frequency ( $P(s)$ ) at all stages of the cell cycle. **f**,  $P(s)$  plots of each individual chromosome (two biological replicates). **g**, A schematic illustrating how compartmentalization levels ( $R(s)$ ) were calculated at different

distance scales (for example, 1 Mb or 100 Mb). Each dotted line indicates a series of 250-kb bin-bin pairs that are separated by a given genomic distance  $s$  (the distance from the diagonal to the dotted line). For all bin-bin pairs separated by distance of  $s$ , a Spearman correlation coefficient  $R(s)$  was generated between observed/expected and the product of two eigenvector 1 values ( $PC1(\text{bin1}) \times PC1(\text{bin2})$ ).  $R(s)$  is expected to be high in well-compartmentalized regions (left) and low at large distance scales with no compartments (right). **h**, Replicate-averaged  $R(s)$  of each individual chromosome across all stages of the cell cycle when  $s$  is equal to 10, 50 and 125 Mb (only eight chromosomes were computed at the 125-Mb scale). **i**, Line graph showing the level of compartmentalization of chr1 against genomic distance at each stage of the cell cycle.



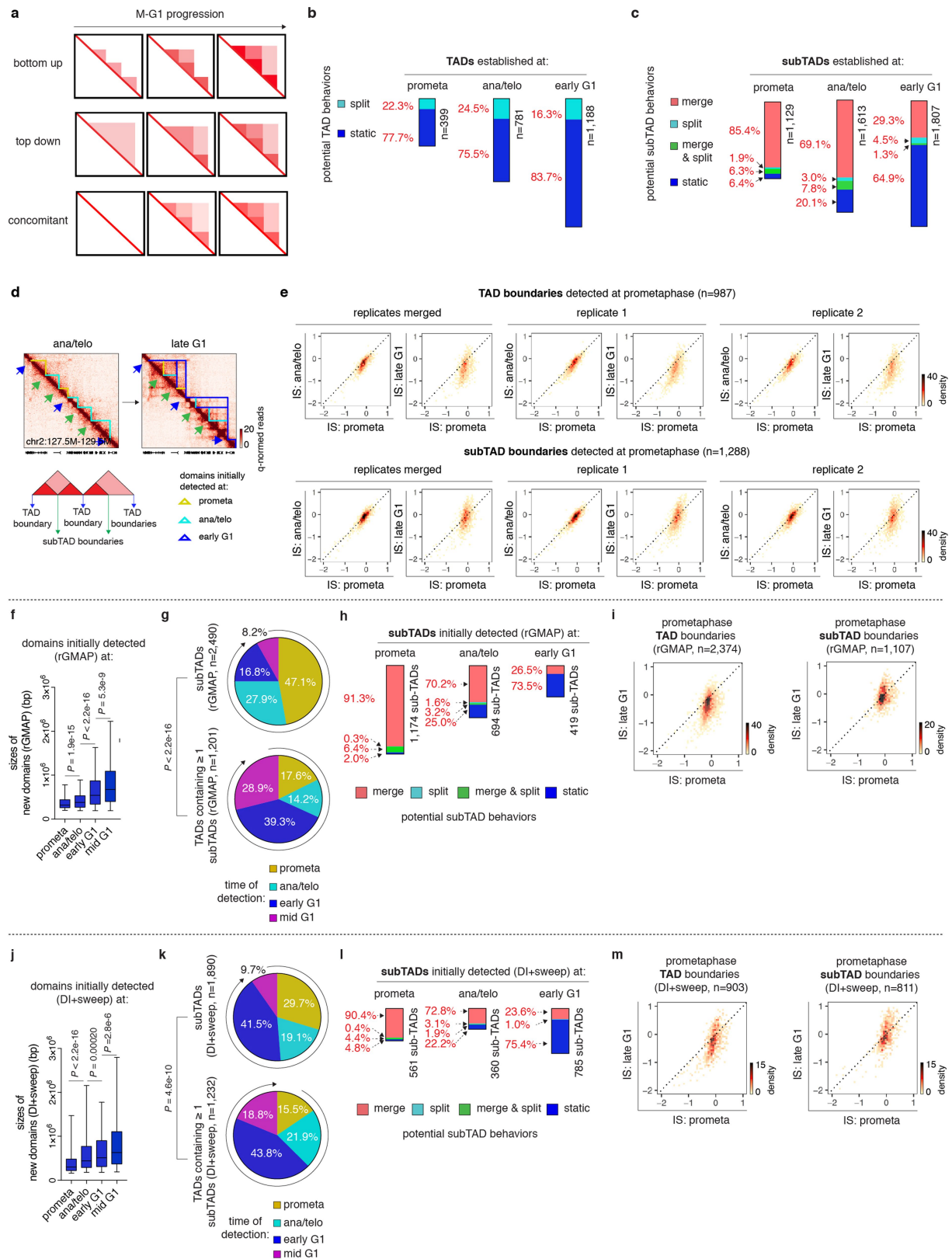


Extended Data Fig. 3 | See next page for caption.



**Extended Data Fig. 3 | Domain detection and residual 'domain-like' structures in prometaphase.** **a, b**, Meta-region plots and density heat maps of insulation scores (**a**) and directionality index (**b**) centred around domain boundaries initially detected at each stage of the cell cycle. **c**, Scatter plots showing Pearson correlations of insulation scores at domain boundaries between two biological replicates. **d**, Aggregated domain analysis (ADA) of domains initially detected at each stage of the cell cycle. **e**, Box plots showing ADA scores over time for domains initially detected at prometaphase ( $n = 1,360$ ), ana/telophase ( $n = 2,260$ ), early G1 ( $n = 2,875$ ) and mid G1 ( $n = 1,112$ ). For all box plots, centre lines denote medians; box limits denote 25th–75th percentile; whiskers denote 5th–95th percentile.  $P$  values were calculated using a two-sided Mann–Whitney  $U$ -test. The dotted line indicates the average ADA score of initial domain detection. **f**, Hi-C contact maps of two representative

regions (chr8: 113 Mb–114 Mb and chr9: 72 Mb–73 Mb) showing residual domain- and boundary-like structures (yellow lines) in prometaphase in merged and individual biological replicates. Bin size, 10 kb. **g**, Simulated featureless, per cent 'G1 contaminated', and early G1 contact maps of the same regions as in **f**. Bin size, 10 kb. **h**, Meta-region plots showing the insulation scores of prometaphase, simulated featureless, 'G1-contaminated' and early G1 samples, centred around prometaphase boundaries in chr8 and chr9. **i**, Meta-region plots showing indicated histone modification profiles centred around boundaries newly detected at each stage of the cell cycle. **j**, Bar graphs showing the enrichment of transcription start sites (overall, housekeeping and tissue-specific<sup>9</sup>) within  $\pm 20$  kb of boundaries newly detected at each stage of the cell cycle.

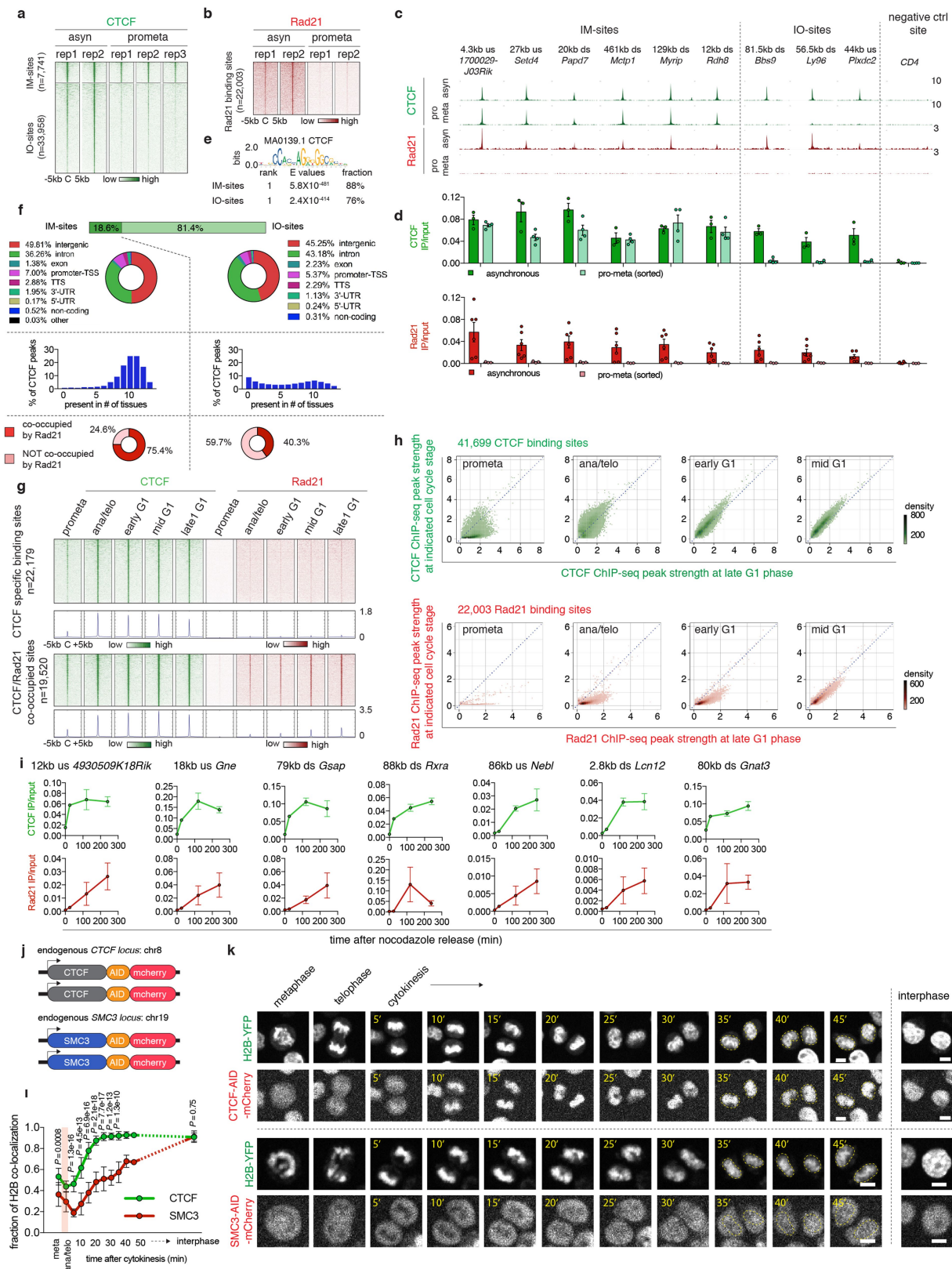


Extended Data Fig. 4 | See next page for caption.

**Extended Data Fig. 4 | Dynamics of TAD and subTAD after mitosis.**

**a**, Schematic of possible models of hierarchical domain formation: bottom-up (merge), top-down (split) and concomitant. **b**, Bar graphs showing the fraction of TADs that display either type of behaviour after detection. **c**, Bar graphs showing the fraction of subTADs that display each of the four potential behaviours after detection: merge, split, merge and split, and static. **d**, Bottom, schematic showing partitioning of boundaries into TAD and subTAD boundaries. Top, Hi-C contact maps showing the change in insulation of representative TAD and subTAD boundaries from ana/telophase to late G1. SubTAD and TAD boundaries are indicated by green and blue arrows, respectively. Bin size, 10 kb. **e**, Bin plots showing the change in insulation score over time of TAD boundaries (top) and subTAD boundaries (bottom) that are detected at prometaphase in merged replicates and in two biological replicates. **f**, Box plots showing sizes of domains initially detected at prometaphase ( $n = 2,494$ ), ana/telophase ( $n = 1,699$ ), early G1 ( $n = 1,357$ ) and mid G1 ( $n = 682$ ) by rGMAP. For all box plots, centre lines denote medians; box limits

denote 25th–75th percentile; whiskers denote 5th–95th percentile. *P* values were calculated using a two-sided Mann–Whitney *U*-test. **g**, Pie charts of the cell cycle distribution of subTADs and TADs that contain at least 1 subTAD based on their time of emergence (called by rGMAP). The *P* value was calculated using a two-sided Fisher's exact test (prometaphase + ana/telophase compared with early G1 + mid G1). **h**, Bar graphs showing the fraction of subTADs detected by rGMAP that display each of the four potential behaviours after detection: merge, split, merge and split, and static. **i**, Bin plots showing the change in insulation score of TAD boundaries (left) and subTAD boundaries (right) that are detected by rGMAP at prometaphase. **j**, Box plots showing the sizes of domains initially detected at prometaphase ( $n = 1,105$ ), ana/telophase ( $n = 1,124$ ), early G1 ( $n = 2,385$ ) and mid G1 ( $n = 520$ ) by DI+sweep (directionality index + window size adjustment). For all box plots, centre lines denote medians; box limits denote 25th–75th percentile; whiskers denote 5th–95th percentile. *P* values were calculated by two-sided Mann–Whitney *U*-test. **k–m**, Similar to **g–i**, showing analyses based on domains called by DI+sweep.

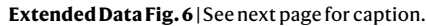


Extended Data Fig. 5 | See next page for caption.

**Extended Data Fig. 5 | CTCF and cohesin chromatin occupancy in mitosis and G1 entry.** **a**, A density heat map of CTCF ChIP-seq data of each biological replicate of asynchronous and prometaphase samples, centred around IM- and IO-CTCF binding sites. **b**, A density heat map of Rad21 ChIP-seq data of both biological replicates of asynchronous and prometaphase samples centred around all Rad21 peaks. **c**, Genome browser tracks showing CTCF and Rad21 ChIP-seq signals of asynchronous and prometaphase samples at indicated regions.  $n = 2-3$  biological replicates. **d**, ChIP-qPCR data of CTCF and Rad21 in asynchronous ( $n = 3$ , 6 biological replicates for CTCF and Rad21, respectively) and prometaphase samples ( $n = 4$ , 3 biological replicates for CTCF and Rad21, respectively). Data are mean  $\pm$  s.e.m. **e**, Motif enrichment analysis of IM- and IO-CTCF binding sites with indicated *E* values as determined by MEME-ChIP. **f**, Top, donut charts showing the genome-wide distribution of IM- and IO-CTCF binding sites. Middle, bar graphs showing the percentage of IM- or IO-CTCF binding sites that are found in indicated numbers of tissues. Bottom, donut pie chart showing the fraction of IM- and IO-CTCF binding sites that are co-occupied by Rad21. **g**, Density heat maps and meta-region plots of CTCF and

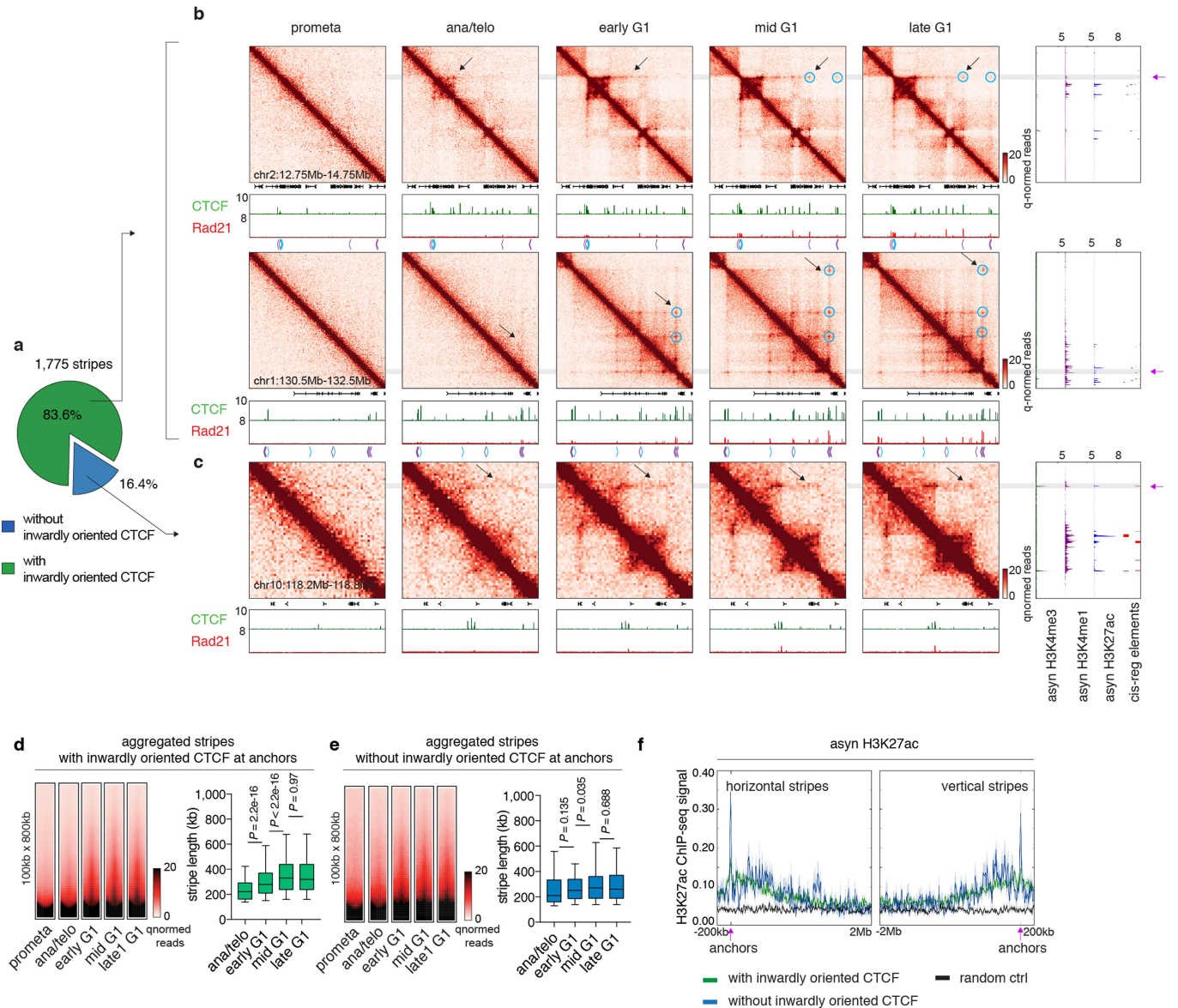
Rad21 ChIP-seq data across all time points centred around CTCF-specific and CTCF/Rad21 co-occupied binding sites. **h**, Bin plots showing ChIP-seq signals of CTCF and Rad21 peaks for each stage of the cell cycle (y axis) against late G1 (x axis). **i**, ChIP-qPCR of CTCF and Rad21 at indicated binding sites over time.  $n = 2$  biological replicates for 0 and 25 min, and  $n = 3$  biological replicates for 120 and 240 min after nocodazole release. Data are mean  $\pm$  s.e.m. **j**, Schematic showing mCherry-tagging of endogenous CTCF and SMC3. **k**, Representative images (from at least 10 dividing cells) illustrating the behaviour of mCherry-tagged CTCF and SMC3 during mitosis-early G1 phase progression. Similar observations were made in 2 independent experiments. Yellow dotted circles demarcate cell nuclei after mitosis. Scale bar, 5  $\mu$ m. **l**, Average recovery curve of mCherry-tagged CTCF and SMC3 that co-localize with H2B-YFP. Cells (11 mother cells/22 daughter cells and 10 mother cells/18 daughter cells) were analysed for CTCF and SMC3, respectively. *P* values were calculated using a two-sided Student's *t*-test. Data are mean  $\pm$  s.e.m. *P* values were omitted at time points with fewer than 5 cells.





**Extended Data Fig. 6 | Loop statistics and *k*-means clustering on structural loops.** **a**, Bar graph showing the number of loop calls at each stage of the cell cycle. **b**, Aggregated peak analysis (APA) of loops initially detected at each stage of the cell cycle. Bin size, 10 kb. Numbers indicate average loop strength:  $\ln(\text{obs}/\text{exp})$ . **c**, Scatter plots showing the Pearson correlation of loop strength (read counts) between two biological replicates. **d**, Hi-C contact maps showing representative regions that contain cluster 1 (chr1: 172.8 Mb–173 Mb), 2 (chr1: 90.2 Mb–90.8 Mb) and 3 (chr2: 47.5 Mb–49 Mb) structural loops in merged and both biological replicates. Bin size, 10 kb. Loop signal enrichment is indicated by black arrows. Contact maps are coupled with genome browser tracks showing CTCF and cohesin occupancy across all stages of the cell cycle. Chevron arrows mark orientations of CTCF sites at loop anchors. **e**, APA of cluster 1, 2 and 3 structural loops across all stages of the cell cycle. Each heat map is coupled with four meta-region plots corresponding to CTCF and Rad21 ChIP-seq signals centred around either upstream or downstream loop anchors. Bin size, 10 kb. Numbers indicate average loop strength:  $\ln(\text{obs}/\text{exp})$ . **f**, Left and right, schematics showing how correlations are computed between CTCF or Rad21 and loop strength over time. Middle, box plot showing the Pearson correlation coefficients between CTCF or Rad21 ChIP-seq peak strength at upstream or downstream anchors and structural loop strength over time ( $n = 4,712$ ). For all box plots, centre lines denote medians; box limits denote 25th–75th percentile; whiskers denote 5th–95th percentile. *P* values

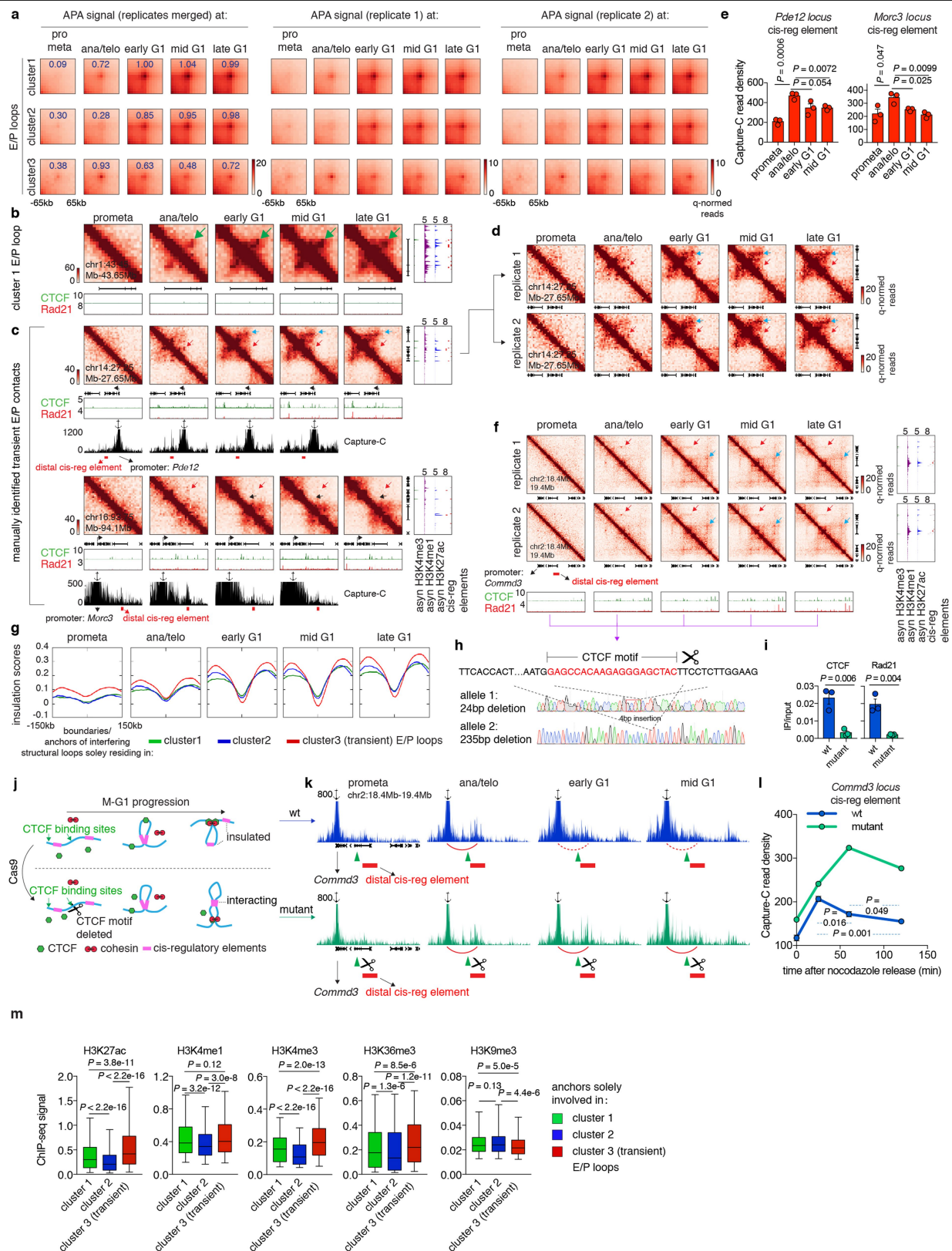
were calculated using a two-sided Wilcoxon signed-rank test. **g**, Box plot showing sizes of structural loops initially detected at ana/telophase ( $n = 90$ ), early G1 ( $n = 2,233$ ), mid G1 ( $n = 1,595$ ) and late G1 ( $n = 793$ ). For all box plots, centre lines denote medians; box limits denote 25th–75th percentile; whiskers denote 5th–95th percentile. *P* values were calculated using a two-sided Mann–Whitney *U*-test. **h**, Average recovery curves of structural loops ( $n = 4,241$ ) and E/P loops with 0 ( $n = 678$ ) or 1 ( $n = 1,338$ ) anchor co-occupied by CTCF/cohesin. The 10% of loops with the smallest increment from prometaphase to late G1 were filtered out from the analysis. Data are mean  $\pm$  99% confidence interval. \*\*\*\* and ####,  $P < 2.2 \times 10^{-16}$  (structural loops compared with E/P loops with 0 or 1 anchor co-occupied by CTCF/cohesin, respectively). Two-sided Mann–Whitney *U*-test. **i**, Left, average recovery curves of randomly sampled and size-matched structural loops and CTCF/cohesin independent E/P loops ( $n = 2,869$  for both groups). The 10% of loops with the smallest increment from prometaphase to late G1 were filtered out from the analysis. Data are mean  $\pm$  99% confidence interval. *P* values were calculated using a two-sided Mann–Whitney *U*-test. Right, box plot showing the comparable size distribution of these two randomly sampled groups ( $n = 2,869$  for both). For both box plots, centre lines denote medians; box limits denote 25th–75th percentile; whiskers denote 5th–95th percentile. **j**, Bar graphs depicting the composition of loops newly called at each stage of the cell cycle.



**Extended Data Fig. 7 | Reformation of chromatin stripes after mitosis. a**, Pie chart showing the fraction of stripes with inwardly oriented CTCF at stripe anchors. **b**, Hi-C contact maps of two representative regions (chr2: 12.75 Mb–14.75 Mb and chr1: 130.5 Mb–132.5 Mb) that contain stripes with inwardly oriented CTCF. Bin size, 10 kb. Contact maps are coupled with genome browser tracks of CTCF and Rad21 across all stages of the cell cycle and tracks of asynchronous H3K4me3, H3K4me1 and H3K27ac and annotation of *cis*-regulatory elements. Chevron arrows mark positions and orientations of CTCF peaks at stripe and loop anchors. Lengthening of stripes is indicated by black arrows. Stripe anchors are indicated by purple arrows. Loops along the stripe axis and at the far end of stripes are indicated by blue circles. **c**, similar to **b**, Hi-C contact maps showing a representative stripe (chr10: 118.2 Mb–118.8 Mb) that

does not have inwardly oriented CTCF at the stripe anchor. **d**, Left, aggregated Hi-C contact maps that compile all stripes with inwardly oriented CTCF to show their overall dynamic growth after mitosis. Right, box plots showing the lengths of these stripes at ana/telophase ( $n = 235$ ), early G1 ( $n = 1,472$ ), mid G1 ( $n = 1,477$ ) and late G1 ( $n = 1,473$ ). For all box plots, centre lines denote medians; box limits denote 25th–75th percentile; whiskers denote 5th–95th percentile. *P* values were calculated using a two-sided Mann–Whitney *U*-test. **e**, Similar to **d**, showing stripes without inwardly oriented CTCF.  $n = 72, 281, 277, 272$  for ana/telophase, early G1, mid G1 and late G1, respectively. **f**, H3K27ac ChIP-seq profile from asynchronous G1E-ER4 cells is plotted –200 kb to 2 Mb around the horizontal stripe anchors and –2 Mb to 200 kb around the vertical stripe anchors. Anchor position is indicated by purple arrows.



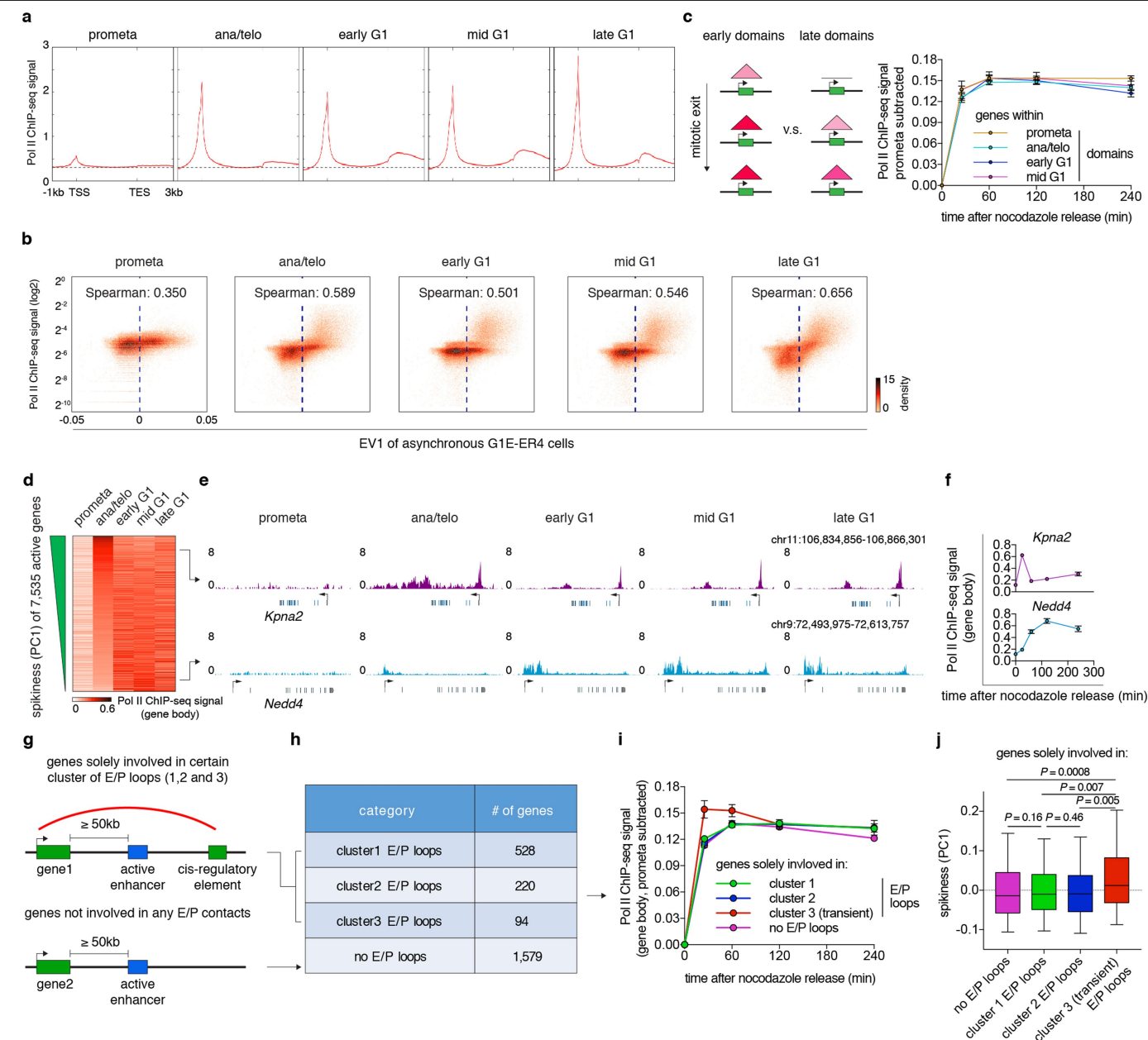


# Article

**Extended Data Fig. 8 | Supplementary E/P loop analyses.** **a**, APA of the three clusters of E/P loops on merged and two biological replicates. Bin size, 10 kb. Numbers indicate average loop strength:  $\ln(\text{obs}/\text{exp})$ . **b**, Hi-C contact maps showing an additional example of cluster 1 E/P loop (chr1: 43.45 Mb–43.65 Mb, green arrow). Bin size, 10 kb. Colour bar denotes  $q$ -normed reads. Contact maps are coupled with genome browser tracks of CTCF and cohesin across all time points as well as asynchronous H3K4me3, H3K4me1 and H3K27ac and annotations of *cis*-regulatory elements. **c**, Similar to **b**, showing two examples of manually identified transient E/P contacts (*Pde12* locus and *Morc3* locus, indicated by red arrow). Boundaries or structural loop anchors that potentially interfere with these E/P contacts are indicated by black and blue arrows, respectively. Contact maps are coupled with tracks of Capture-C interaction profiles. Probes (anchor symbol) are located at promoters of *Pde12* and *Morc3* genes. **d**, Hi-C contact maps showing the *Pde12* locus on two biological replicates. Bin size, 10 kb. **e**, Quantification of the Capture-C read density of the red regions in **c**.  $n = 3$  biological replicates. Data are mean  $\pm$  s.e.m.  $P$  values were calculated from two-sided Student's  $t$ -test. **f**, Similar to **d**, Hi-C contact maps showing the cluster3 E/P loop (red arrows) at *Commd3* locus in two biological replicates. Potential interfering loop is indicated by blue arrows. **g**, Insulation score profiles centred around the boundaries and interfering structural loop anchors that solely reside within cluster 1, 2 or 3 E/P loops. **h**, Sanger

sequencing profiles showing deletion of the CTCF core motif at the upstream anchor of the structural loop (blue arrows in **f**) that potentially interfere with the cluster3 E/P loop at the *Commd3* locus (red arrows in **f**). **i**, ChIP–qPCR showing the abrogation of CTCF and Rad21 binding at the edited site in **f**.  $n = 3$  biological replicates. Data are mean  $\pm$  s.e.m.  $P$  values were calculated by two-sided Student's  $t$ -test. **j**, Schematic showing potential behaviour of cluster 3 E/P loops before and after deletion of the interfering structural loop anchor. **k**, Capture-C interaction profiles between *Commd3* promoter and downstream *cis*-regulatory element (red bars) on wild-type and interfering anchor-deleted mutant cells over time. The location of the capture probe is indicated by the anchor symbol. The deleted CTCF site is indicated by green triangles. Formation of the transient loop is indicated by red arches. **l**, Quantification showing read density of the red regions in **k**.  $n = 3$  and 2 biological replicates for wild-type and mutant cells, respectively. Data are mean  $\pm$  s.e.m.  $P$  values were calculated by two-sided Student's  $t$ -test. **m**, Box plots showing ChIP–seq signals of indicated histone modifications at anchors that solely participate in cluster 1, 2 or 3 (transient) E/P loops ( $n = 2,612, 1,338$  and 413 respectively). For all box plots, centre lines denote medians; box limits denote 25th–75th percentile; whiskers denote 5th–95th percentile.  $P$  values were calculated using a two-sided Mann–Whitney  $U$ -test.





**Extended Data Fig. 9 | Relationship between post-mitotic structural organization and gene reactivation. a**, Meta-region analysis of Pol II occupancy of active genes across all stages of the cell cycle. TSS, transcription start site; TES, transcription end site. **b**, Bin plots showing the positive correlation between Pol II ChIP-seq signal strength and eigenvector 1 (asynchronous G1E-ER4 cells<sup>27</sup>, 25-kb binned) genome-wide. **c**, Left, schematic showing genes that are within early or late domains. Right, average Pol II occupancy of genes that reside in prometaphase ( $n = 2,274$  genes) ana/telophase ( $n = 2,114$  genes), early G1 ( $n = 1,159$  genes) and mid G1 ( $n = 303$  genes) emerging domains. Data are mean  $\pm$  99% confidence interval. **d**, Heat map showing gene-body Pol II occupancy across all stages of the cell cycle. Genes are ranked by their PC1 values ('spikiness'). **e**, Genome browser tracks showing

representative examples of early spiking (*Kpna2*) and gradually activating (*Nedd4*) genes. **f**, Quantification of gene-body Pol II occupancy in **e**.  $n = 2$  biological replicates for 0 h, and  $n = 3$  biological replicates for other time points. Data are mean  $\pm$  s.e.m. **g**, Schematic showing the stratification of genes on the basis of their involvement in E/P loops. **h**, Table showing the number of genes that are solely involved in clusters of E/P loops. **i**, Average gene-body Pol II occupancy of the genes in **h** over time. Sample sizes are shown in **h**. Data are mean  $\pm$  s.e.m. **j**, Box plots showing the spikiness (PC1) of genes in **h**. Sample sizes are shown in **h**. For all box plots, centre lines denote medians; box limits denote 25th–75th percentile; whiskers denote 5th–95th percentile.  $P$  values were calculated using a two-sided Mann–Whitney  $U$ -test.

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- |                                     |   |
|-------------------------------------|---|
| n/a                                 | Confirmed   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of all covariates tested   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. $F$ , $t$ , $r$ ) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted<br><i>Give <math>P</math> values as exact values whenever suitable.</i>                 |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> Estimates of effect sizes (e.g. Cohen's $d$ , Pearson's $r$ ), indicating how they were calculated  |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

### Software and code

Policy information about [availability of computer code](#)

#### Data collection

For flow data collection, we used FACSDiva 8 (BD Biosciences) and Summit (Beckman Coulter)  
For image collection, we used MetaMorph  
For high-throughput sequencing data collection, we used NextSeq Control Software v2.2.0.

#### Data analysis

For statistical analyses, we used R (R studio) or GraphPad Prism 7.  
For imaging processing, we used Fiji (Image J 2.0.0).  
For flow chart generation, we used FlowJo 10.4.0.  
For high throughput sequencing data processing and subsequent data analyses we used:  
pandas 0.22.0,  
scipy 0.19.1,  
numpy 1.13.3,  
HiC-Pro\_2.7.7,  
juicer\_tools\_0.7.5.jar, 0.7.0 jar  
FastQC 0.11.5,  
trim galore 0.4.1-0,  
Cutadapt 1.18,  
FLASH 1.2.8,  
CCAnalyzer3,  
bowtie 2  
bowtie 0.12.7,  
SAMtools v0.1.19,  
macs2 v2.1.0, 2.1.1,  
kent UCSC Utilities,  
bwtool 1.0,  
HOMER 4.9.1,

BEDtools, 2.27.1

deeptools 2.5.4

For loop and stripe identification and DI+sweep. See method section (code is available upon request to the corresponding authors)

3D netmod domain calling: [https://bitbucket.org/creminslab/3dnetmod\\_method\\_v3.0\\_development](https://bitbucket.org/creminslab/3dnetmod_method_v3.0_development)

rGMAP 1.4

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All figures include publicly available data. All ChIP-seq, Capture-C and Hi-C raw and processed data generated from this study are now deposited into the GEO database with accession number GSE129997 for public access. ChIP-seq files of histone modifications shown in figure 4 and Extended data figures 7 and 8 are available from the GEO database with accession numbers: H3K27ac (GSE61349), H3K4me1 (GSM946535), H3K4me3 (GSM946533), H3K36me3 (GSM946529) and H3K9me3 (GSM946542)

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was not pre-determined. We used sample sizes commonly accepted for high throughput genome wide experiments. We performed 2 biological replicates for Hi-C, 2-3 biological replicates for ChIP-seq and ChIP-qPCR, and 2-3 biological replicates for Capture-C. Hi-C and ChIP-seq data were pooled for down-stream analyses.
Data exclusions	Experiments were done in multiple replicates. Replicates with technical failure were removed.
Replication	2 biological replicates for Hi-C and at least 2-3 biological replicates for ChIP-seq and capture-C were generated.
Randomization	Experiments were not randomized. No animal or human subjects were involved in this study.
Blinding	Researchers were not blind to group allocation. Blinding was not relevant to our study as no human subjects were involved.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Antibodies

Antibodies used	anti-pMPM2 Millipore, catlog#: 05-368, Clone: MPM-2, multiple lots of antibodies were used. Dilution: 0.2ul/10million cells anti-lamin B1 Abcam, catlog#: ab16048, Polyclonal, lot# GR-3214420-1. Dilution: 0.5ul/1million cells
-----------------	---

anti-CTCF Millipore, catlog#: 07-729, Polyclonal, lot# 2887267. Dilution: 5-10ug/ChIP

anti-Rad21 Abcam, catlog#: ab992, Polyclonal, lot# GR214359-7. Dilution: 5-10ug/ChIP

anti-Pol II Cell Signaling, catlog#: 14958, Clone: D8L4Y, lot# 1. Dilution: 8ug/IP

F(ab')<sub>2</sub>-goat anti-mouse secondary antibody, APC, Thermo Fisher Scientific, catelog#: 17-4010-82. Polyclonal, lot# 1997054. Dilution: 20ul/10million cells.

Fluorescein (FITC) AffiniPure Goat Anti-Rabbit IgG (H+L), Jackson ImmunoResearch, Code: 111-095-144, Polyclonal. Dilution: 0.5ul/1million cells

## Validation

anti-pMPM2 Millipore, catlog#: 05-368, multiple lots of antibodies were used. This antibody has been claimed to react with mouse pMPM2 by the manufacturer.

anti-lamin B1 Abcam, catlog#: ab16048, lot# GR-3214420-1. This antibody has been claimed to react with mouse by the manufacturer. This antibody has been extensively used to assess lamin B1 distribution in the literature (eg. Klymenko et al. Leukemia. 2018). Our experiments also confirmed the correct nuclear peripheral distribution of lamin B1 in interphase cells using this antibody.

anti-CTCF Millipore, catlog#: 07-729, lot# 2887267. This antibody has been claimed to react with mouse and validated for ChIP-seq by the manufacturer. This antibody was also extensively used for ChIP-seq studies and confirmed to lose signal upon CTCF depletion in our hands and also by others (eg. Nora. et al. 2017)

anti-Rad21 Abcam, catlog#: ab992, lot# GR214359-7. This antibody has been claimed to react with mouse and validated for ChIP experiments by the manufacturer. This antibody was also confirmed to lose signal upon loss of Rad21 (Rao. et al. 2017).

anti-Pol II Cell Signaling, 14958, lot# 1. This antibody has been claimed to react with mouse and to be suitable for ChIP by the manufacture. This antibody has also been previously used in our lab (Behera et al. 2019, Cell Reports) and also by others for ChIP experiments (eg. Sun Y. et al. 2019, Science Advances.)

F(ab')<sub>2</sub>-goat anti-mouse secondary antibody, APC, Thermo Fisher Scientific, catelog#: 17-4010-82. lot# 1997054. The manufacturer has tested this antibody to be suitable for immunofluorescence studies.

Fluorescein (FITC) AffiniPure Goat Anti-Rabbit IgG (H+L), Jackson ImmunoResearch, Code: 111-095-144. the manufacturer states that this antibody reacts with whole molecule rabbit IgG. This antibody has been previously used by other studies on mouse samples (eg. Li et al. eLife. 2018). We also confirmed that it's positive for immunofluorescence.

## Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)

The G1E-ER4 cell line was a gift from Mitchell Weiss' laboratory.

Authentication

We regularly confirm that these cells can be induced to undergo terminal erythroid differentiation.

Mycoplasma contamination

G1E-ER4 cells has been tested to be negative of Mycoplasma

Commonly misidentified lines  
(See [ICLAC](#) register)

The cell line used (G1E-ER4) is not in the ICLAC database

## ChIP-seq

### Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links

*May remain private before publication.*

The Hi-C, Capture-C and ChIP-seq data generated and analyzed in this study are deposited in GEO repository under accession number GSE129997 for public access.

Files in database submission

1899\_1900-1\_S1\_Read1.fq.gz  
1899\_1900-1\_S1\_Read2.fq.gz  
1899\_1900-2\_S1\_Read1.fq.gz  
1899\_1900-2\_S1\_Read2.fq.gz  
1899\_1900-3\_S1\_Read1.fq.gz  
1899\_1900-3\_S1\_Read2.fq.gz  
1899\_1900-4\_S1\_Read1.fq.gz  
1899\_1900-4\_S1\_Read2.fq.gz  
1901-1902-5\_S1\_Read1.fq.gz  
1901-1902-5\_S1\_Read2.fq.gz  
1901-1902-6\_S1\_Read1.fq.gz

1901-1902-6\_S1\_Read2.fq.gz  
1901-1902-7\_S1\_Read1.fq.gz  
1901-1902-7\_S1\_Read2.fq.gz  
1901-1902-8\_S1\_Read1.fq.gz  
1901-1902-8\_S1\_Read2.fq.gz  
1903-1904-10\_S1\_Read1.fq.gz  
1903-1904-10\_S1\_Read2.fq.gz  
1903-1904-11\_S1\_Read1.fq.gz  
1903-1904-11\_S1\_Read2.fq.gz  
1903-1904-12\_S1\_Read1.fq.gz  
1903-1904-12\_S1\_Read2.fq.gz  
1903-1904-9\_S1\_Read1.fq.gz  
1903-1904-9\_S1\_Read2.fq.gz  
1900\_1\_S2\_Read1.fq.gz  
1900\_1\_S2\_Read2.fq.gz  
1900\_2\_S2\_Read1.fq.gz  
1900\_2\_S2\_Read2.fq.gz  
1900\_3\_S2\_Read1.fq.gz  
1900\_3\_S2\_Read2.fq.gz  
1900\_4\_S2\_Read1.fq.gz  
1900\_4\_S2\_Read2.fq.gz  
1902\_5\_S2\_Read1.fq.gz  
1902\_5\_S2\_Read2.fq.gz  
1902\_6\_S2\_Read1.fq.gz  
1902\_6\_S2\_Read2.fq.gz  
1902\_7\_S2\_Read1.fq.gz  
1902\_7\_S2\_Read2.fq.gz  
1902\_8\_S2\_Read1.fq.gz  
1902\_8\_S2\_Read2.fq.gz  
1904\_9\_S2\_Read1.fq.gz  
1904\_9\_S2\_Read2.fq.gz  
1904\_10\_S2\_Read1.fq.gz  
1904\_10\_S2\_Read2.fq.gz  
1904\_11\_S2\_Read1.fq.gz  
1904\_11\_S2\_Read2.fq.gz  
1904\_12\_S2\_Read1.fq.gz  
1904\_12\_S2\_Read2.fq.gz  
1530\_run151\_Read1.fastq.gz  
1530\_run151\_Read2.fastq.gz  
1573\_run163\_Read1.fastq.gz  
1573\_run163\_Read2.fastq.gz  
1573\_run165\_Read1.fastq.gz  
1573\_run165\_Read2.fastq.gz  
1554\_run154\_Read1.fastq.gz  
1554\_run154\_Read2.fastq.gz  
1554\_run160\_Read1.fastq.gz  
1554\_run160\_Read2.fastq.gz  
1554\_run161\_Read1.fastq.gz  
1554\_run161\_Read2.fastq.gz  
1554\_run162\_Read1.fastq.gz  
1554\_run162\_Read2.fastq.gz  
1531\_run151\_Read1.fastq.gz  
1531\_run151\_Read2.fastq.gz  
1574\_run163\_Read1.fastq.gz  
1574\_run163\_Read2.fastq.gz  
1574\_run165\_Read1.fastq.gz  
1574\_run165\_Read2.fastq.gz  
1555\_run154\_Read1.fastq.gz  
1555\_run154\_Read2.fastq.gz  
1555\_run160\_Read1.fastq.gz  
1555\_run160\_Read2.fastq.gz  
1555\_run161\_Read1.fastq.gz  
1555\_run161\_Read2.fastq.gz  
1555\_run162\_Read1.fastq.gz  
1555\_run162\_Read2.fastq.gz  
1536\_run152\_Read1.fastq.gz  
1536\_run152\_Read2.fastq.gz  
1536\_run155\_Read1.fastq.gz  
1536\_run155\_Read2.fastq.gz  
1536\_run156\_Read1.fastq.gz  
1536\_run156\_Read2.fastq.gz  
1556\_run154\_Read1.fastq.gz  
1556\_run154\_Read2.fastq.gz  
1556\_run160\_Read1.fastq.gz  
1556\_run160\_Read2.fastq.gz



1556\_run161\_Read1.fastq.gz  
1556\_run161\_Read2.fastq.gz  
1556\_run162\_Read1.fastq.gz  
1556\_run162\_Read2.fastq.gz  
1537\_run152\_Read1.fastq.gz  
1537\_run152\_Read2.fastq.gz  
1537\_run155\_Read1.fastq.gz  
1537\_run155\_Read2.fastq.gz  
1537\_run156\_Read1.fastq.gz  
1537\_run156\_Read2.fastq.gz  
1557\_run154\_Read1.fastq.gz  
1557\_run154\_Read2.fastq.gz  
1557\_run160\_Read1.fastq.gz  
1557\_run160\_Read2.fastq.gz  
1557\_run161\_Read1.fastq.gz  
1557\_run161\_Read2.fastq.gz  
1557\_run162\_Read1.fastq.gz  
1557\_run162\_Read2.fastq.gz  
1533\_run151\_Read1.fastq.gz  
1533\_run151\_Read2.fastq.gz  
1533\_run155\_Read1.fastq.gz  
1533\_run155\_Read2.fastq.gz  
1533\_run156\_Read1.fastq.gz  
1533\_run156\_Read2.fastq.gz  
1558\_run154\_Read1.fastq.gz  
1558\_run154\_Read2.fastq.gz  
1558\_run160\_Read1.fastq.gz  
1558\_run160\_Read2.fastq.gz  
1558\_run161\_Read1.fastq.gz  
1558\_run161\_Read2.fastq.gz  
1558\_run162\_Read1.fastq.gz  
1558\_run162\_Read2.fastq.gz  
1930\_S1\_Read1.fq.gz  
1930\_S1\_Read2.fq.gz  
1931\_S1\_Read1.fq.gz  
1931\_S1\_Read2.fq.gz  
1932\_S1\_Read1.fq.gz  
1932\_S1\_Read2.fq.gz  
1933\_S1\_Read1.fq.gz  
1933\_S1\_Read2.fq.gz  
1934\_S1\_Read1.fq.gz  
1934\_S1\_Read2.fq.gz  
1935\_S1\_Read1.fq.gz  
1935\_S1\_Read2.fq.gz  
1936\_S1\_Read1.fq.gz  
1936\_S1\_Read2.fq.gz  
1937\_S1\_Read1.fq.gz  
1937\_S1\_Read2.fq.gz  
1930\_S2\_Read1.fq.gz  
1930\_S2\_Read2.fq.gz  
1931\_S2\_Read1.fq.gz  
1931\_S2\_Read2.fq.gz  
1932\_S2\_Read1.fq.gz  
1932\_S2\_Read2.fq.gz  
1933\_S2\_Read1.fq.gz  
1933\_S2\_Read2.fq.gz  
1934\_S2\_Read1.fq.gz  
1934\_S2\_Read2.fq.gz  
1935\_S2\_Read1.fq.gz  
1935\_S2\_Read2.fq.gz  
1936\_S2\_Read1.fq.gz  
1936\_S2\_Read2.fq.gz  
1937\_S2\_Read1.fq.gz  
1937\_S2\_Read2.fq.gz  
CTCF\_asyn\_Rep0.fastq  
CTCF\_asyn\_Rep1.fastq  
CTCF\_prometa\_Rep0.fastq  
CTCF\_prometa\_Rep1.fastq  
CTCF\_prometa\_Rep2.fastq  
CTCF\_anatelo\_Rep5.fastq  
CTCF\_anatelo\_Rep6.fastq  
CTCF\_earlyG1\_Rep1.fastq  
CTCF\_earlyG1\_Rep2.fastq  
CTCF\_midG1\_Rep1.fastq  
CTCF\_midG1\_Rep2.fastq

CTCF\_lateG1\_Rep1.fastq  
 CTCF\_lateG1\_Rep2.fastq  
 Rad21\_asyn\_Rep1.fastq  
 Rad21\_asyn\_rep3.fastq  
 Rad21\_prometa\_Rep3.fastq  
 Rad21\_prometa\_Rep4.fastq  
 Rad21\_anatelo\_Rep5.fastq  
 Rad21\_anatelo\_Rep6.fastq  
 Rad21\_earlyG1\_Rep1.fastq  
 Rad21\_earlyG1\_Rep5.fastq  
 Rad21\_midG1\_Rep1.fastq  
 Rad21\_midG1\_Rep3.fastq  
 Rad21\_lateG1\_Rep1.fastq  
 Rad21\_lateG1\_Rep3.fastq  
 wt\_prometa\_rep1\_Read1.fq.gz  
 wt\_prometa\_rep2\_Read1.fq.gz  
 wt\_ana.telo\_rep1\_Read1.fq.gz  
 wt\_ana.telo\_rep2\_Read1.fq.gz  
 wt\_ana.telo\_rep3\_Read1.fq.gz  
 wt\_early\_g1\_rep1\_Read1.fq.gz  
 wt\_early\_g1\_rep2\_Read1.fq.gz  
 wt\_early\_g1\_rep3\_Read1.fq.gz  
 wt\_mid\_g1\_rep1\_Read1.fq.gz  
 wt\_mid\_g1\_rep2\_Read1.fq.gz  
 wt\_mid\_g1\_rep3\_Read1.fq.gz  
 wt\_late\_g1\_rep1\_Read1.fq.gz  
 wt\_late\_g1\_rep2\_Read1.fq.gz  
 wt\_late\_g1\_rep3\_Read1.fq.gz  
 input\_asyn.fastq  
 input\_prometa.fastq  
 input\_anatelo.fastq  
 input\_earlyG1.fastq  
 input\_midG1.fastq  
 input\_lateG1.fastq  
 capture\_c\_commd3\_locus\_wt\_pro\_meta\_merged.bdg  
 capture\_c\_commd3\_locus\_wt\_ana.telo\_merged.bdg  
 capture\_c\_commd3\_locus\_wt\_early\_G1\_merged.bdg  
 capture\_c\_commd3\_locus\_wt\_mid\_G1\_merged.bdg  
 capture\_c\_morc3\_locus\_wt\_pro\_meta\_merged.bdg  
 capture\_c\_morc3\_locus\_wt\_ana.telo\_merged.bdg  
 capture\_c\_morc3\_locus\_wt\_early\_G1\_merged.bdg  
 capture\_c\_morc3\_locus\_wt\_mid\_G1\_merged.bdg  
 capture\_c\_pde12\_locus\_wt\_pro\_meta\_merged.bdg  
 capture\_c\_pde12\_locus\_wt\_ana.telo\_merged.bdg  
 capture\_c\_pde12\_locus\_wt\_early\_G1\_merged.bdg  
 capture\_c\_pde12\_locus\_wt\_mid\_G1\_merged.bdg  
 capture\_c\_slow\_structural\_loop\_wt\_pro\_meta\_merged.bedgraph  
 capture\_c\_slow\_structural\_loop\_wt\_ana.telo\_merged.bedgraph  
 capture\_c\_slow\_structural\_loop\_wt\_early\_G1\_merged.bedgraph  
 capture\_c\_slow\_structural\_loop\_wt\_mid\_G1\_merged.bedgraph  
 capture\_c\_fast\_structural\_loop\_wt\_pro\_meta\_merged.bedgraph  
 capture\_c\_fast\_structural\_loop\_wt\_ana.telo\_merged.bedgraph  
 capture\_c\_fast\_structural\_loop\_wt\_early\_G1\_merged.bedgraph  
 capture\_c\_fast\_structural\_loop\_wt\_mid\_G1\_merged.bedgraph  
 prometa\_raw\_intrachromosomal\_contact\_matrices.tar.gz  
 prometa\_kr\_intrachromosomal\_contact\_matrices.tar.gz  
 prometa\_qnorm\_intrachromosomal\_contact\_matrices.tar.gz  
 anatelo\_raw\_intrachromosomal\_contact\_matrices.tar.gz  
 anatelo\_kr\_intrachromosomal\_contact\_matrices.tar.gz  
 anatelo\_qnorm\_intrachromosomal\_contact\_matrices.tar.gz  
 earlyG1\_raw\_intrachromosomal\_contact\_matrices.tar.gz  
 earlyG1\_kr\_intrachromosomal\_contact\_matrices.tar.gz  
 earlyG1\_qnorm\_intrachromosomal\_contact\_matrices.tar.gz  
 midG1\_raw\_intrachromosomal\_contact\_matrices.tar.gz  
 midG1\_kr\_intrachromosomal\_contact\_matrices.tar.gz  
 midG1\_qnorm\_intrachromosomal\_contact\_matrices.tar.gz  
 lateG1\_raw\_intrachromosomal\_contact\_matrices.tar.gz  
 lateG1\_kr\_intrachromosomal\_contact\_matrices.tar.gz  
 lateG1\_qnorm\_intrachromosomal\_contact\_matrices.tar.gz  
 prometa\_domains.bed  
 anatelo\_domains.bed  
 earlyG1\_domains.bed  
 midG1\_domains.bed  
 lateG1\_domains.bed  
 prometa\_loops.tsv

```

anatelo_loops.tsv
earlyG1_loops.tsv
midG1_loops.tsv
lateG1_loops.tsv
prometa_stripes.tsv
anatelo_stripes.tsv
earlyG1_stripes.tsv
midG1_stripes.tsv
lateG1_stripes.tsv
CTCF_asyn.bw
CTCF_prometa.bw
CTCF_anatelo.bw
CTCF_earlyG1.bw
CTCF_midG1.bw
CTCF_lateG1.bw
Rad21_asyn.bw
Rad21_prometa.bw
Rad21_anatelo.bw
Rad21_earlyG1.bw
Rad21_midG1.bw
Rad21_lateG1.bw
CTCF_asyn.narrowPeak
CTCF_prometa.narrowPeak
CTCF_anatelo.narrowPeak
CTCF_earlyG1.narrowPeak
CTCF_midG1.narrowPeak
CTCF_lateG1.narrowPeak
Rad21_asyn.narrowPeak
Rad21_prometa.narrowPeak
Rad21_anatelo.narrowPeak
Rad21_earlyG1.narrowPeak
Rad21_midG1.narrowPeak
Rad21_lateG1.narrowPeak
capture_c_commd3_locus_commd3_mutant_prometa_merged.bdg
capture_c_commd3_locus_commd3_mutant_anateo_merged.bdg
capture_c_commd3_locus_commd3_mutant_early_g1_merged.bdg
capture_c_commd3_locus_commd3_mutant_mid_g1_merged.bdg
WT_Pol2_Noc0h.bw
WT_Pol2_Noc25min.bw
WT_Pol2_Noc1h.bw
WT_Pol2_Noc2h.bw
WT_Pol2_Noc4h.bw

```

Genome browser session  
(e.g. [UCSC](https://genome.ucsc.edu/s/thomasgilgenast/mitosis-ctcf-rad21-pol2-chipseq))

<https://genome.ucsc.edu/s/thomasgilgenast/mitosis-ctcf-rad21-pol2-chipseq>

## Methodology

### Replicates

See methods section. Briefly, 2-3 biological replicates were generated per time point per antibody. Specifically:

For CTCF, we generated 2 biological replicates per time point except prometaphase, for which we performed 3 biological replicates.

For Rad21, we performed 2 biological replicates per time point.

For Pol II, we performed 3 biological replicates per time point except prometaphase, for which we performed 2 biological replicates.

### Sequencing depth

See supplementary table 7

### Antibodies

anti-CTCF Millipore, 07-729  
anti-Rad21 Abcam, ab992  
anti-Pol II Cell Signaling, 14958

### Peak calling parameters

Sequencing reads were mapped to the reference mouse genome mm9 using bowtie (0.12.7, "-m 2 --tryhard"). Reads were filtered to remove non-uniquely mapped reads and PCR duplicates using Samtools (v0.1.19) and converted to bed format using BEDtools (v2.27.1, "bedtools bamtobed").

For CTCF and Rad21, filtered reads from each biological replicate were pooled together and down-sampled to equivalent read counts across all cell cycle stages. Peaks were identified using the MACS2 with punctate calling for both CTCF and Rad21 (p-values  $1e-8$  and  $1e-4$  respectively), using each IP's cell cycle stage matched input as the control.

For Pol II, Peaks were then called with MACS2 for each replicate with a p-value cutoff of  $1e-4$ , using each IP's cell cycle stage matched input as the control.

## Data quality

- (1). Raw fastq files were assessed with FastQC (v0.11.5) prior to processing.
- (2). Peaks were called using input controls corresponding to every cell cycle stage.
- (3). Peaks were called using p-value cutoffs described above and the default above 5-fold enrichment.
- (4). Correlation among replicates was assessed (Extended Data Fig. 1g-i).
- (5). Peaks of Rad21 and CTCF were largely overlapping and motif analysis revealed the expected CTCF motif within peaks.
- (6). Peaks of Pol II were largely located at expected genomic regions (TSS, gene bodies).

## Software

For ChIP-seq data processing and analyses, we used bowtie version 0.12.7, Samtools v0.1.19, macs2 2.1.1, 2.1.0, kent UCSC Utilities, bwtool version 1.0, HOMER version 4.9.1, BEDtools and deepTools 2.5.4

## Flow Cytometry

### Plots

Confirm that:

- ☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☒ All plots are contour plots with outliers or pseudocolor plots.
- ☐ A numerical value for number of cells or percentage (with statistics) is provided.

### Methodology

## Sample preparation

See method section.

Briefly, actively proliferating G1E-ER4 cells were synchronized at pro-metaphase with nocodazole treatment. Cells were then released from nocodazole for several durations (0h, 25min, 1h, 2h and 4h). Cells were harvested, washed with PBS and crosslinked with 1% PFA for 10min. Crosslinks were quenched with glycine for 5min and cells were then permeabilized with 0.1% Triton X-100. Finally, cells were stained with 20ng/ml DAPI and subjected to cell sorting.

## Instrument

Beckman Coulter MoFlo Astrios sorter/Becton Dickinson FACS Aria Fusion sorter

## Software

Flow charts were generated using FlowJo 10.4.0

## Cell population abundance

We achieved very high cell purity of pro-metaphase populations. The purity of pro-metaphase cells was >98%. This was confirmed by DAPI staining and the disassembly of lamin B1.

## Gating strategy

Prometaphase cells were gated on mcherry (high), DAPI (4N) and pMPM2 (high) fluorescent signal. ana/telophase cells were gated based on DAPI (4N) and mcherry (low, relative to prometaphase) signals. G1 samples were gated based on DAPI (2N) and mcherry (low) signals.

- ☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

# The structures and gating mechanism of human calcium homeostasis modulator 2

<https://doi.org/10.1038/s41586-019-1781-3>

Wooyoung Choi<sup>1,4</sup>, Nicolina Clemente<sup>1,4</sup>, Weinan Sun<sup>2,3</sup>, Juan Du<sup>1\*</sup> & Wei Lü<sup>1\*</sup>

Received: 8 March 2019

Accepted: 8 October 2019

Published online: 27 November 2019

Calcium homeostasis modulators (CALHMs) are voltage-gated,  $\text{Ca}^{2+}$ -inhibited nonselective ion channels that act as major ATP release channels, and have important roles in gustatory signalling and neuronal toxicity<sup>1–3</sup>. Dysfunction of CALHMs has previously been linked to neurological disorders<sup>1</sup>. Here we present cryo-electron microscopy structures of the human CALHM2 channel in the  $\text{Ca}^{2+}$ -free active or open state and in the ruthenium red (RUR)-bound inhibited state, at resolutions up to 2.7 Å. Our work shows that purified CALHM2 channels form both gap junctions and undecameric hemichannels. The protomer shows a mirrored arrangement of the transmembrane domains (helices S1–S4) relative to other channels with a similar topology, such as connexins, innexins and volume-regulated anion channels<sup>4–8</sup>. Upon binding to RUR, we observed a contracted pore with notable conformational changes of the pore-lining helix S1, which swings nearly 60° towards the pore axis from a vertical to a lifted position. We propose a two-section gating mechanism in which the S1 helix coarsely adjusts, and the N-terminal helix fine-tunes, the pore size. We identified a RUR-binding site near helix S1 that may stabilize this helix in the lifted conformation, giving rise to channel inhibition. Our work elaborates on the principles of CALHM2 channel architecture and symmetry, and the mechanism that underlies channel inhibition.

ATP release channels have a fundamental role in many neurological functions—including the modulation of excitatory synaptic strength, long-term synaptic potentiation and neuronal excitability—by mediating the purinergic signalling pathway in the central nervous system<sup>9,10</sup>. CALHMs act as one of the major ATP release channels, together with maxi-anion channels, volume-regulated anion channels (VRACs), connexins and pannexins<sup>4–8,11</sup>. CALHMs are abundantly expressed in taste bud cells and have an important role in sensing sweet, bitter and umami flavours<sup>3</sup>. They are activated by a reduction in extracellular calcium and membrane depolarization, which triggers a signalling cascade in the neural gustatory pathways<sup>12</sup>. CALHMs also have an important role in cortical neuron excitability<sup>13</sup>. Dysregulation of CALHM1, as well as its P86L polymorphism, have previously been linked to neurological disorders such as Alzheimer's disease and ischaemic brain damage<sup>1</sup>.

CALHMs are predicted to have four transmembrane helices, similar to connexins, innexins, pannexins and VRACs. However, the architecture, symmetry and domain arrangement of CALHMs remain unknown. Connexins and VRACs are hexamers<sup>4–7</sup> and innexins are octamers<sup>8</sup>; the current concept is that CALHM1 and CALHM3 form hexamers and that they do not form gap junctions, owing to *N*-glycosylation in the extracellular loop<sup>11,14</sup>.

In addition to  $\text{Ca}^{2+}$ , CALHMs can be modulated by a variety of small molecules that includes RUR,  $\text{Gd}^{3+}$  and 2-aminoethoxydiphenyl borate (2-APB)<sup>2,12,14–17</sup>. RUR is a hexavalent polysaccharide stain<sup>18</sup> that non-specifically inhibits many ion channels<sup>19–21</sup> through an unknown molecular mechanism.

## Structural determination

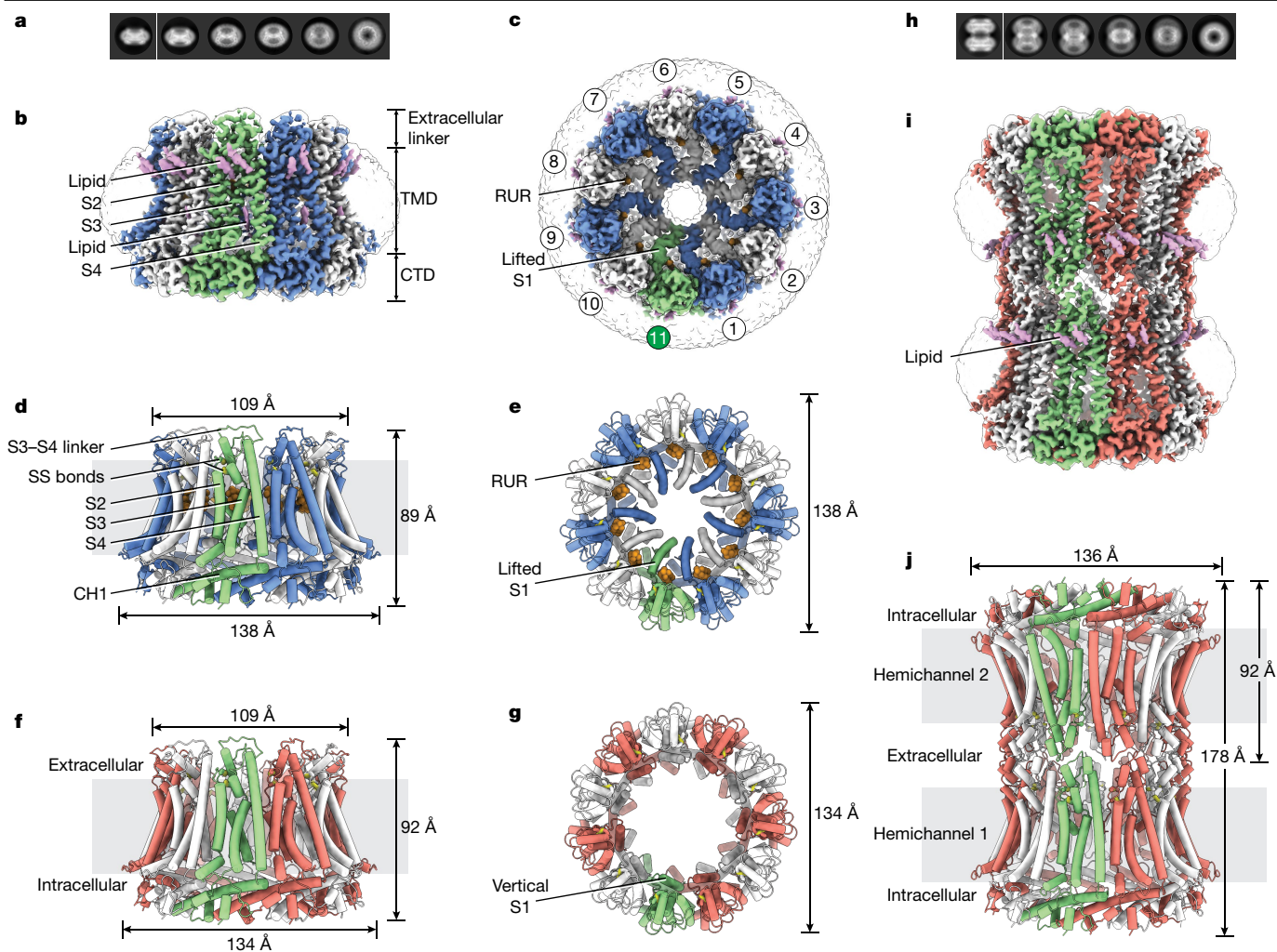
Our whole-cell electrophysiology data showed that human CALHM2 produces a robust current in the absence of  $\text{Ca}^{2+}$ . The current showed no obvious voltage dependence but was inhibited by  $\text{Ca}^{2+}$  or RUR in a voltage-dependent manner (Extended Data Fig. 1a, c–f). To elucidate the assembly of CALHM2 and to understand the molecular mechanism that underlies channel gating, we studied the human CALHM2 channel in the presence of EDTA or the antagonist RUR using cryo-electron microscopy (cryo-EM).

The initial cryo-EM experiments using GFP-tagged constructs yielded solely hemichannel particles. The two-dimensional classes show a fuzzy tail, which is probably the GFP tag, and the three-dimensional reconstructions were of low resolution (Extended Data Fig. 2a, c, d). The GFP-cleaved construct not only improved the cryo-EM map but also showed both hemichannel and gap-junction particles at a ratio of approximately 1:1 at grid concentration (about 18  $\mu\text{M}$ ), representing the approximate dissociation constant ( $K_d$ ) of the docking of hemichannels (Extended Data Figs. 2b, 3). Unlike the CALHM1 channel<sup>14</sup>, CALHM2 did not show *N*-glycosylation (Extended Data Fig. 2e, f), which probably explains the existence of a gap junction. Only hemichannels, and not gap junctions, were observed in the presence of RUR (Extended Data Fig. 4). These observations indicate that both the C-terminal GFP tag and the binding of RUR may affect the conformation of the docking site.

We determined the structures of CALHM2 in the presence of EDTA as hemichannels and gap junctions (EDTA–CALHM2<sup>hemi</sup> and

<sup>1</sup>Van Andel Institute, Grand Rapids, MI, USA. <sup>2</sup>Vollum Institute, Oregon Health & Science University, Portland, OR, USA. <sup>3</sup>Present address: Janelia Research Campus, Ashburn, VA, USA. <sup>4</sup>These authors contributed equally: Wooyoung Choi, Nicolina Clemente. \*e-mail: [juan.du@vai.org](mailto:juan.du@vai.org); [wei.lu@vai.org](mailto:wei.lu@vai.org)





**Fig. 1 | The overall architecture of CALHM2.** Odd-numbered subunits are in blue (for RUR–CALHM2) or red (for EDTA–CALHM2<sup>hemi</sup>), and even-numbered subunits are in white. The eleventh subunit is in green, lipid-like densities are in purple and RUR densities are in orange. **a**, Selected two-dimensional class averages of RUR–CALHM2. **b**, **c**, The three-dimensional reconstruction of RUR–CALHM2, viewed parallel to the membrane (**b**) and from extracellular side

of the membrane (**c**). SS bonds, disulfide bonds. **d–g**, The structures of RUR–CALHM2 (**d**, **e**) and EDTA–CALHM2<sup>hemi</sup> (**f**, **g**). **h**, Selected two-dimensional class averages of EDTA–CALHM2<sup>gap</sup>. **i**, **j**, The three-dimensional reconstruction (**i**) and atomic model (**j**) of EDTA–CALHM2<sup>gap</sup>. Unsharpened reconstructions are shown as transparent envelopes.

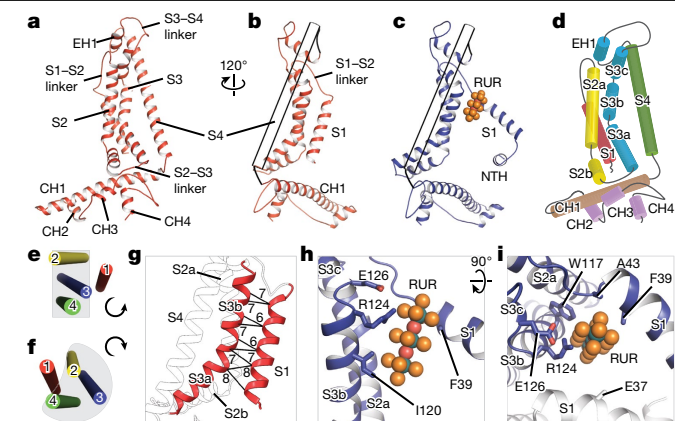
EDTA–CALHM2<sup>gap</sup>, respectively) and in the presence of RUR (RUR–CALHM2), at resolutions of 3.3, 3.5 and 2.7 Å, respectively (Extended Data Figs. 3–6, Supplementary Table 1). Our structures unambiguously showed that CALHM2 is an undecamer (Fig. 1a–g). Moreover, we observed 11 strong cylinder-shaped densities in the RUR–CALHM2 structure, which represent RUR molecules (Fig. 1c, e, Extended Data Fig. 4).

The S1 helix and the N-terminal helix (NTH) were poorly defined relative to the rest of the channel, indicating high flexibility. To assess the conformational heterogeneity of these helices, we applied an approach that combined symmetry expansion and signal subtraction, in which all of the subunits were subtracted and classified<sup>22</sup> (Extended Data Figs. 3, 4). In RUR–CALHM2, two distinct conformations of helix S1 appeared, with 70% in the lifted conformation and 13% in the vertical conformation; the rest were not well-defined. RUR densities were observed only in the class with lifted helix S1. By contrast, EDTA–CALHM2<sup>hemi</sup> contained approximately half of the helix S1 in a vertical conformation, and the rest was invisible. There are non-resolvable densities in the pore vestibule that may represent the intermediate states of helix S1 between the lifted and vertical conformations.

We use these two extreme conformations—RUR–CALHM2 with all of helix S1 in the lifted conformation and EDTA–CALHM2 with all S1 in the vertical conformation—to discuss the gating mechanism and the action of the antagonist RUR (Fig. 1c, e, g). Owing to intrinsic positional uncertainty, we fitted only the protein backbone into the densities of the S1 helix and NTH, and the exact position of residues in this region (residues 13–40) should be interpreted with caution. Nevertheless, we provide electrophysiology data that validate the placement of helix S1 (discussed in the section ‘Subunit structure and RUR-binding site’). A part of the extracellular loops is also poorly defined (residues 138–152).

## Overall architecture

The CALHM2 structure assembles as an undecamer with each protomer consisting of a large N-terminal transmembrane domain (TMD) (which comprises helices S1, S2, S3 and S4, and NTH), an intracellular C-terminal domain (CTD) and a small extracellular linker region (Fig. 1b–g). The hemichannels show an overall truncated cone shape that consists of helices S2, S3 and S4 as a side wall, the CTD as a base and the



**Fig. 2 | A single subunit of CALHM2, and Rur-binding site.** **a–c**, Cartoon representation of EDTA–CALHM2<sup>hemi</sup> (**a**, **b**) and Rur–CALHM2 (**c**). The S4 helix in **b** and **c** is shown as a transparent tube for clarity. **d**, Domain organization of EDTA–CALHM2<sup>hemi</sup>. **e**, **f**, TMD organization of EDTA–CALHM2<sup>hemi</sup> (**e**) and connexin 43 (**f**) viewed from the extracellular side. **g**, The loose contacts between the S1 and S3 helices in EDTA–CALHM2<sup>hemi</sup>. Distances (in Å) between Cα of adjacent residues in helices S1 and S3 are labelled. **h**, **i**, Rur-binding site.

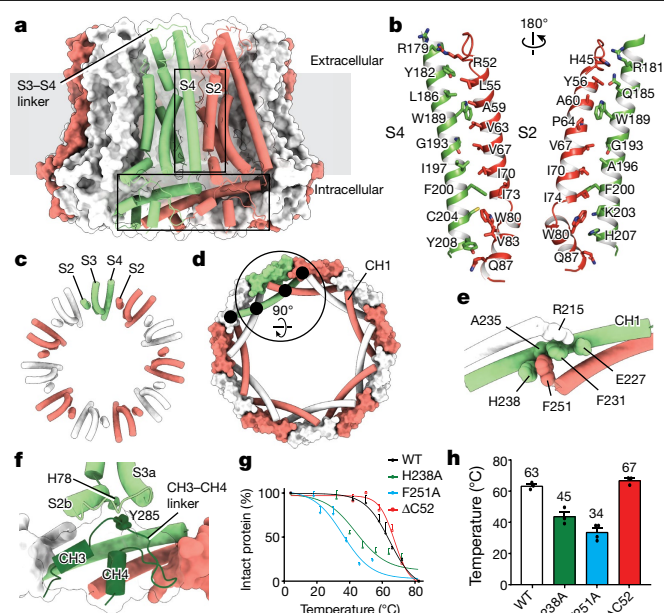
extracellular linker region as a rim, with the intracellular side wider than the extracellular. The first long helix in the CTD (CH1) intercrosses neighbouring subunits, and thus weaves CALHM2 into a sturdy undecamer with a diameter that is 1.6- and 1.2-fold larger than those of connexins and innexins, respectively (Extended Data Fig. 7).

When viewed from the extracellular side, the EDTA–CALHM2<sup>hemi</sup> structure contains an unusually large pore that is lined by helix S1, which is poised parallel to the pore axis; we thus term this conformation ‘vertical S1’ (Fig. 1g). By comparison, Rur–CALHM2 shows a vertical compression and a horizontal expansion (Fig. 1d–g). Most notably, the pore in Rur–CALHM2 is markedly smaller: helix S1 swings towards the pore axis in what we term the ‘lifted S1’ conformation (Fig. 1c, e). A Rur density was observed underneath helix S1, and probably supports this helix in the lifted conformation—thus stabilizing the channel in an inhibited state. This agrees with previous functional studies that have shown that Rur inhibits CALHMs<sup>2,12,14–17</sup> (Extended Data Fig. 1a, c). Notably, the intracellular halves of helices S3 and S4 are relatively far apart, which creates a gap in which a lipid-like density is observed (Fig. 1b). Such a loose contact between the S3 and S4 helices might facilitate the conformational change of S3 during channel gating because the pore-lining S1 is attached to S3.

Similar to connexins and innexins, EDTA–CALHM2<sup>gap</sup> is docked by two hemichannels in a head-to-head manner, forming a thick cylinder with a large diameter (Fig. 1h–j). The docking region of EDTA–CALHM2<sup>gap</sup> is considerably shorter than the gap junctions of connexin and innexin, owing to a smaller extracellular linker region (Extended Data Fig. 7a, c, d). Two disulfide bonds connect the S3–S4 linker and S1–S2 linker, which may have an important role in stabilizing the docking of two hemichannels<sup>23</sup> (Fig. 1d). Disulfide bonds are present in similar positions in connexins and innexins<sup>4,7,8,23</sup>. Finally, we observed two strong lipid-like densities close to the extracellular linker region, which probably contribute to maintaining the integrity of the docking site (Fig. 1b, i).

### Subunit structure and Rur-binding site

The protomer of CALHM2 is L-shaped: two long and straight helices—helix S4 in the TMD and CH1 in the CTD—run nearly perpendicular to each other (Fig. 2a–c). Both the S2 and S3 helices are multi-segment helices. The S2a and S3b segments form a plane together with helix S4, which runs parallel to the pore axis; the S2b and S3a segments reside on

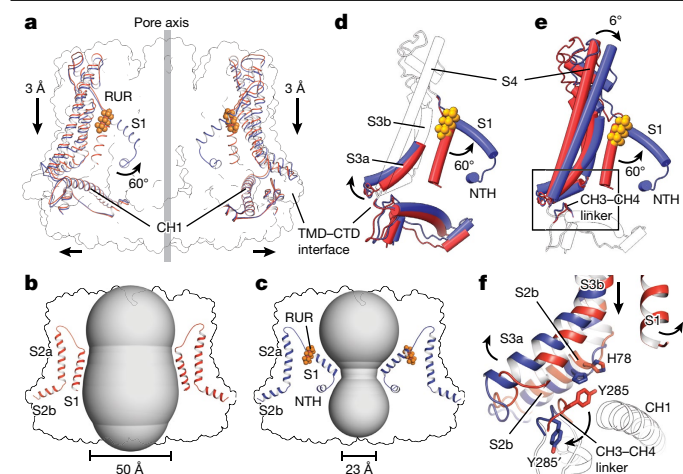


**Fig. 3 | Inter- and intrasubunit interactions.** **a**, A view of the intersubunit interface of EDTA–CALHM2<sup>hemi</sup> at the TMD (vertical rectangle; shown in **b**, **c**) and at the CTD (horizontal rectangle). **b**, **c**, The intersubunit interface at the TMD between helix S2 and the adjacent S4, viewed in parallel to the membrane (**b**) or from the intracellular side (**c**). **d**, Interface at the CTD viewed from the intracellular side. Black filled circles indicate the four intersubunit interfaces of CH1. **e**, Enlargement of the large circle in **d** showing the interactions of three neighbouring CH1 helices. **f**, Enlargement of the large circle in **d**, showing the TMD–CTD interface and the exterior circular layer in the CTD formed by CH3, CH4 and the CH3–CH4 linker. **g**, **h**, Thermodynamic test of wild-type CALHM2, the mutants of key residues involved in the CH1 interactions in **e**, and the mutant with a deletion of S2 residues at the C-terminal end (ΔC52). Both F251A and H238A considerably decreased the thermostability of the protein, but the ΔC52 construct showed no effect.  $n = 3$ , 3, 4 or 3 biologically independent experiments were performed for wild-type CALHM2, CALHM2(H238A), CALHM2(F251A) or CALHM2(ΔC52), respectively. Data are mean  $\pm$  s.e.m. Each dot indicates the value of one single independent experiment.

the top of the CTD, mediating the only contact between the TMD and the CTD (Fig. 2d). A comparison of the CALHM2 protomer with those of connexins, innexins and VRACs showed notable differences in size, shape and domain organization (Extended Data Fig. 7c, d).

The most notable distinction between CALHM2 and other channels with a similar topology (such as connexins, innexins and VRACs) is at the TMD. Viewed from the extracellular side, the S1, S2, S3 and S4 helices are arranged anticlockwise in EDTA–CALHM2<sup>hemi</sup>, with S2, S3 and S4 approximately in a plane; by contrast, the arrangements in connexins, innexins and VRACs are all clockwise, with helices S2, S3 and S4 forming a compact helix bundle (Fig. 2e, f). As a consequence, the S1 helix in EDTA–CALHM2<sup>hemi</sup> is loosely attached only to the S3 (Fig. 2e, g), but in connexins, innexins and VRACs, it is clamped in the helix bundle and forms extensive interactions with S2 and S4 (Fig. 2f). We suggest that such a loose contact of helix S1 to the rest of the TMD in CALHM2 gives this helix a high flexibility to swing up and down. In Rur–CALHM2, helix S1 was detached from S3, and a Rur molecule occupied the vertical S1 position (Fig. 2c, h, i). Because Rur is positively charged, to validate the Rur-binding site and the placement of the lifted S1, we studied a charge-reversing mutant of E37 (a key residue on helix S1 that interacts with Rur). Although CALHM2(E37R) displays Ca<sup>2+</sup>-dependent gating similar to that of wild-type CALHM2, the inhibition of this mutant by Rur is nearly abolished (Extended Data Fig. 1b, c). Underneath the lifted S1, we also observed part of the NTH, which is involved in voltage-sensing in CALHM1, connexin, pannexin and innexin<sup>7,24–28</sup> (Fig. 2c).





**Fig. 4 | RUR inhibition and ion-conducting pore.** **a**, Superimposition of EDTA-CALHM2<sup>hemi</sup> (red) and RUR-CALHM2 (blue) using the 11 CH1 helices. Only two subunits are shown, for clarity. **b**, **c**, The shape of the ion-conducting pore in EDTA-CALHM2<sup>hemi</sup> (**b**) and in RUR-CALHM2 (**c**). The smallest pore diameters were estimated without considering the involvement of the NTHs; they do not represent the real pore size and are used only to show the change of the pore profile. **d**, **e**, Superimposition of a single subunit of EDTA-CALHM2<sup>hemi</sup> (red) and RUR-CALHM2 (blue) using either the TMD (**d**) or the CTD (**e**). Regions with conformational changes are highlighted in colour and the rest is shown as a transparent cartoon, for clarity. **f**, Enlargement of the rectangle in **e**, showing conformational changes between EDTA-CALHM2<sup>hemi</sup> (red) and RUR-CALHM2 (blue) at the TMD-CTD interface. The CTD is shown as a transparent cartoon for clarity.

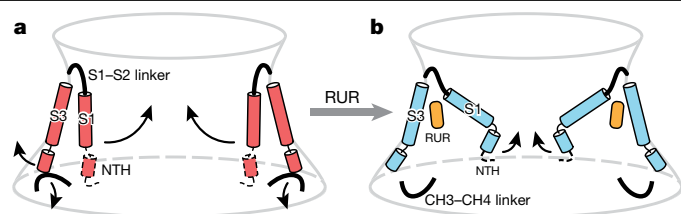
The CTD in CALHM2 is the only component on the intracellular side and has an extended shape due to the long CH1, plus the three short  $\alpha$ -helices (CH2, CH3 and CH4) (Fig. 2a–d). By comparison, the intracellular regions of innexins and VRACs are formed by both the CTD and the S2–S3 linker (Extended Data Fig. 7c, d). The extracellular side of CALHM2 consists mainly of a flat S3–S4 linker that is involved in the docking of two hemichannels (Fig. 2a). The S1–S2 linker, on the other hand, is very short and not directly involved in the docking region. The longer S1–S2 linker in connexin and innexin protrudes into the extracellular space, forms an intact structure with the S3–S4 linker and participates directly in the docking (Extended Data Fig. 7c, d).

### Channel assembly

The CALHM2 hemichannel exhibits a compact exterior through extensive subunit–subunit interactions with three major interfaces, one at the TMD between adjacent subunits (Fig. 3a–c), and the other two at the CTD (Fig. 3d–f). The extracellular linker region of CALHM2 lacks interactions, which presumably provides flexibility for the docking of two hemichannels into a gap junction (Fig. 3a).

At the TMD layer, only the S2 and S4 helices of adjacent subunits interact with each other, mainly through hydrophobic interactions (Fig. 3b, c). By contrast, the TMD in connexins, innexins and VRACs forms a compact helix bundle, which results in different inter-subunit interfaces. Specifically, where in connexin the S1 and S2 helices of adjacent subunits mediate the inter-subunit interface<sup>4</sup>, in innexin there is a lack of interaction between adjacent TMDs<sup>8</sup>.

At the CTD layer, each long CH1 projects out to interact with 4 adjacent CH1 helices (2 on each side), weaving the 11 subunits into a large circular frame (Fig. 3d). We identified a major interaction located in the middle of CH1, sandwiched by the N- and C-terminal regions of two adjacent CH1 helices. These three adjacent CH1 helices dovetail



**Fig. 5 | Schematic of the RUR-induced inhibition mechanism.** Conformational changes are shown between EDTA-CALHM2<sup>hemi</sup> (**a**) and RUR-CALHM2 (**b**). The observed movement of helix S1 and the proposed movement of the NTH during channel inhibition are indicated. The region that cannot be modelled is indicated by dashed lines.

to each other through aromatic and charged residues (Fig. 3e), which are conserved in the CALHM family and have crucial roles in channel assembly and stability (Fig. 3g, h, Extended Data Fig. 8). Moreover, CH3 and CH4—along with their linker—buckle into the space of the circular frame, forming an additional exterior circular layer (Fig. 3d, f). Such a complex interaction network in the CTD is absent in connexins, innexins and VRACs. The contact between the TMD and the CTD is mediated mainly by the CH3–CH4 linker with its cognate S2b and S3a segments, through the interaction between Y285 and H78 (Fig. 3f). Notably, a P86L polymorphism in CALHM1 that is linked with Alzheimer’s disease is located in the TMD–CTD interface (Extended Data Fig. 8). This polymorphism changes the functional properties of CALHM1<sup>1,2,15,29</sup>, which suggests this interface has an important role in channel gating.

### Inhibition mechanism by RUR

We compared the structures of EDTA-CALHM2<sup>hemi</sup> and RUR-CALHM2 (Fig. 4, Supplementary Video 1). The 11 CH1 helices of each structure are well-aligned, which provides support for the role of CH1 in constituting a rigid scaffold of the channel. By contrast, the TMD displays notable conformational changes in two aspects. First, RUR occupies the position of the vertical S1, and thus drives helix S1 to swing nearly 60° towards the pore axis; this reduces the pore diameter by approximately 27 Å (Fig. 4b, c). Second, there is obvious vertical compression and horizontal expansion, resulting in the entire TMD sliding towards the CTD and remodelling at the TMD–CTD interface (Fig. 4a).

To further understand the action of RUR, we compared single subunits of EDTA-CALHM2<sup>hemi</sup> and RUR-CALHM2 by superimposing their CTD or TMD (Fig. 4d, e). With the exceptions of helix S1 and the TMD–CTD interface (Fig. 4d), the TMD displays a rigid-body inward tilting towards the pore axis upon binding of RUR, which explains the vertical compression of the channel (Fig. 4e). In EDTA-CALHM2<sup>hemi</sup>, the vertical S1 loosely attaches to helix S3, with the S3a segment sterically restricting the CH3–CH4 linker on the CTD. The CTD is further coupled to the S2b segment through the interaction between Y285 on the CH3–CH4 linker and H78 on the S2b, thus supporting the TMD in an elevated position (Fig. 4f). The binding of RUR detaches helix S1 from S3, and segment S3a swaps outward and releases its restriction on the CH3–CH4 linker. As a result, Y285 flips nearly 180° (Fig. 4f), rupturing the coupling between Y285 and H78 and leading the entire TMD to move towards the CTD.

To define the functional states of EDTA-CALHM2<sup>hemi</sup> and RUR-CALHM2, we inspected their ion-conducting pores. The pore diameters of EDTA-CALHM2 and RUR-CALHM2 at two extreme conditions, with all S1 in the vertical or lifted conformation, were estimated to be 50 and 23 Å (respectively) without considering the 11 copies of the NTH (Fig. 4b, c). Although the NTHs are disordered in the high-resolution structures refined using  $C_{11}$  symmetry, we observed prominent S1 helix and NTH densities restricting the pore in one of the asymmetric

RUR–CALHM2 classes (Extended Data Fig. 4), implying that the real pore sizes should be smaller. Indeed, truncation of the first 20 N-terminal residues (CALHM2(ΔN20)) showed a notable reduction in inhibition by RUR, which provides support for a role of the NTH in channel gating by physically restricting the pore; such a role is consistent with other channels with a similar topology (including connexins<sup>4,7</sup>, innexins<sup>8</sup> and VRACs<sup>5,6</sup>). Moreover, a charge-neutralizing R10A mutant (CALHM2(R10A)) in the NTH showed a marked reduction in inhibition by RUR at negative membrane potentials, indicating the involvement of the NTH in the voltage-dependent inhibition by RUR (Extended Data Figs. 1a, d–f, 9).

We suggest that the EDTA–CALHM2<sup>hemi</sup> structure represents an active or open state, and the RUR–CALHM2 structure represents an inhibited state. The RUR functions as an antagonist instead of a pore blocker, based on its binding site. Despite being a nonspecific inhibitor for many ion channels<sup>19–21</sup>, this is—to our knowledge—the first time that the molecular mechanism that underlies inhibition by RUR has been elucidated. Relative to wild-type CALHM2, Ca<sup>2+</sup>-dependent inhibition remained unaffected for CALHM2(ΔN20) and CALHM2(R10A), implying that RUR and Ca<sup>2+</sup> may not share the same binding site and that they may inhibit CALHM2 through different mechanisms (Extended Data Figs. 1a, d–f, 9).

## CALHM2 gap junction

The EDTA–CALHM2 gap junction is docked by two hemichannels through the extracellular S3–S4 linker (Extended Data Fig. 10a). In contrast to the hemichannel, the pore in EDTA–CALHM2<sup>gap</sup> is smaller, and helix S1 is in a lifted conformation similar to that in RUR–CALHM2 (Extended Data Fig. 10b). The S1 helix is less well-defined than those in RUR–CALHM2 because it lacks the stabilization of the RUR molecule underneath. The docking of two hemichannels results in a marked conformational rearrangement at the junction, in which a flat loop in the S3–S4 linker is reformed into a triangle shape. As a result, two interfaces are created at the junction (Extended Data Fig. 10c, d), a primary interface between the paired subunits of two hemichannels and a minor interface between the diagonal subunits. The deletion of a segment (Δ143–146) in the S3–S4 linker hindered CALHM2 in forming a gap junction (Extended Data Fig. 10d, e).

Further investigation is required to reveal whether EDTA–CALHM2<sup>gap</sup> is in an inhibited state similar to that of RUR–CALHM2, and how the two hemichannels within a gap junction coordinate with each other. In addition, despite the observation of a CALHM2 gap junction in vitro, its existence in vivo remains to be determined.

## Conclusion

Our CALHM2 structures showed the undecameric assembly of a CALHM family member, and describe a molecular mechanism that underlies inhibitory gating induced by the antagonist RUR (Fig. 5). The primary determinants for the channel gate of CALHM2 are the S1 helix and NTH. The Ca<sup>2+</sup>-free hemichannel favours helix S1 in a vertical conformation and loosely attached to S3, resulting in a large open pore. When RUR occupies the space in which the vertical S1 is located, it stabilizes helix S1 upward, which contracts the pore. We propose a two-section inhibitory mechanism, in which the S1 helix adjusts the pore size coarsely and the NTH makes fine adjustments, eventually physically occluding the pore. We speculate that helix S1 and the NTH of the I1 subunits may move individually (rather than concertedly) to assume the conformational change and thus adjust the pore size with high flexibility, which enables molecules of various sizes to pass through the pore. Our structures build a solid foundation for understanding the physiology and pharmacology of the CALHM family.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1781-3>.

- Dreses-Werringloer, U. et al. A polymorphism in CALHM1 influences Ca<sup>2+</sup> homeostasis, Aβ levels, and Alzheimer's disease risk. *Cell* **133**, 1149–1161 (2008).
- Ma, Z. et al. Calcium homeostasis modulator 1 (CALHM1) is the pore-forming subunit of an ion channel that mediates extracellular Ca<sup>2+</sup> regulation of neuronal excitability. *Proc. Natl Acad. Sci. USA* **109**, E1963–E1971 (2012).
- Taruno, A. et al. CALHM1 ion channel mediates purinergic neurotransmission of sweet, bitter and umami tastes. *Nature* **495**, 223–226 (2013).
- Maeda, S. et al. Structure of the connexin 26 gap junction channel at 3.5 Å resolution. *Nature* **458**, 597–602 (2009).
- Deneke, D., Sawicka, M., Lam, A. K. M., Paulino, C. & Dutzler, R. Structure of a volume-regulated anion channel of the LRRC8 family. *Nature* **558**, 254–259 (2018).
- Kefauver, J. M. et al. Structure of the human volume regulated anion channel. *eLife* **7**, e38461 (2018).
- Myers, J. B. et al. Structure of native lens connexin 46/50 intercellular channels by cryo-EM. *Nature* **564**, 372–377 (2018).
- Oshima, A., Tani, K. & Fujiyoshi, Y. Atomic structure of the innexin-6 gap junction channel determined by cryo-EM. *Nat. Commun.* **7**, 13681 (2016).
- Abbracchio, M. P., Burnstock, G., Verkhratsky, A. & Zimmermann, H. Purinergic signalling in the nervous system: an overview. *Trends Neurosci.* **32**, 19–29 (2009).
- Burnstock, G. Historical review: ATP as a neurotransmitter. *Trends Pharmacol. Sci.* **27**, 166–176 (2006).
- Ma, Z. et al. CALHM3 is essential for rapid ion channel-mediated purinergic neurotransmission of GPCR-mediated tastes. *Neuron* **98**, 547–561 (2018).
- Ma, Z., Tanis, J. E., Taruno, A. & Foskett, J. K. Calcium homeostasis modulator (CALHM) ion channels. *Pflügers Arch.* **468**, 395–403 (2016).
- Ma, Z., Saung, W. T. & Foskett, J. K. Action potentials and ion conductances in wild-type and CALHM1-knockout type II taste cells. *J. Neurophysiol.* **117**, 1865–1876 (2017).
- Siebert, A. P. et al. Structural and functional similarities of calcium homeostasis modulator 1 (CALHM1) ion channel with connexins, pannexins, and innexins. *J. Biol. Chem.* **288**, 6140–6153 (2013).
- Dreses-Werringloer, U. et al. CALHM1 controls the Ca<sup>2+</sup>-dependent MEK, ERK, RSK and MSK signaling cascade in neurons. *J. Cell Sci.* **126**, 1199–1206 (2013).
- Vingtreux, V. et al. CALHM1 ion channel elicits amyloid-β clearance by insulin-degrading enzyme in cell lines and in vivo in the mouse brain. *J. Cell Sci.* **128**, 2330–2338 (2015).
- Cisneros-Mejorado, A. et al. Blockade and knock-out of CALHM1 channels attenuate ischemic brain damage. *J. Cereb. Blood Flow Metab.* **38**, 1060–1069 (2018).
- Martinez-Palomo, A., Baislovsky, C. & Bernhard, W. Ultrastructural modifications of the cell surface and intercellular contacts of some transformed cell strains. *Cancer Res.* **29**, 925–937 (1969).
- Caterina, M. J., Rosen, T. A., Tominaga, M., Brake, A. J. & Julius, D. A capsaicin-receptor homologue with a high threshold for noxious heat. *Nature* **398**, 436–441 (1999).
- Kirichok, Y., Navarro, B. & Clapham, D. E. Whole-cell patch-clamp measurements of spermatozoa reveal an alkaline-activated Ca<sup>2+</sup> channel. *Nature* **439**, 737–740 (2006).
- Peier, A. M. et al. A heat-sensitive TRP channel expressed in keratinocytes. *Science* **296**, 2046–2049 (2002).
- Bai, X. C., Rajendra, E., Yang, G., Shi, Y. & Scheres, S. H. Sampling the conformational space of the catalytic subunit of human γ-secretase. *eLife* **4**, e11182 (2015).
- Foot, C. I., Zhou, L., Zhu, X. & Nicholson, B. J. The pattern of disulfide linkages in the extracellular loop regions of connexin 32 suggests a model for the docking interface of gap junctions. *J. Cell Biol.* **140**, 1187–1197 (1998).
- Purnick, P. E., Oh, S., Abrams, C. K., Verselis, V. K. & Bargiello, T. A. Reversal of the gating polarity of gap junctions by negative charge substitutions in the N-terminus of connexin 32. *Biophys. J.* **79**, 2403–2415 (2000).
- Verselis, V. K., Ginter, C. S. & Bargiello, T. A. Opposite voltage gating polarities of two closely related connexins. *Nature* **368**, 348–351 (1994).
- Oh, S., Rivkin, S., Tang, Q., Verselis, V. K. & Bargiello, T. A. Determinants of gating polarity of a connexin 32 hemichannel. *Biophys. J.* **87**, 912–928 (2004).
- Oh, S., Abrams, C. K., Verselis, V. K. & Bargiello, T. A. Stoichiometry of transjunctional voltage-gating polarity reversal by a negative charge substitution in the amino terminus of a connexin32 chimera. *J. Gen. Physiol.* **116**, 13–31 (2000).
- Michalski, K., Henze, E., Nguyen, P., Lynch, P. & Kawate, T. The weak voltage dependence of pannexin 1 channels can be tuned by N-terminal modifications. *J. Gen. Physiol.* **150**, 1758–1768 (2018).
- Moreno-Ortega, A. J., Ruiz-Núñez, A., García, A. G. & Cano-Abad, M. F. Mitochondria sense with different kinetics the calcium entering into HeLa cells through calcium channels CALHM1 and mutated P86L-CALHM1. *Biochem. Biophys. Res. Commun.* **391**, 722–726 (2010).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

## Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

### Cloning

The full-length human *CALHM1*, *CALHM2* and *CALHM3* genes (<http://www.uniprot.org>; UniProtKB numbers: Q8IU99, Q9HA72 and Q86XJ0, respectively) were synthesized by Genscript and subcloned into a pEG BacMam vector comprising an N- or C-terminal thrombin cleavage site, GFP and His8 tag<sup>30</sup>. We focused on CALHM2 because it showed the best biochemical properties among CALHM1, CALHM2 and CALHM3 (Extended Data Fig. 1g–i). CALHM2-mutant primers were synthesized from Eurofins USA and were produced using site-directed mutagenesis and/or primer extension PCR.

### Construct screening

tsA201 cells were seeded into 2-ml plates to a final density of  $0.9\text{--}1 \times 10^6$  cells/ml in DMEM containing 10% (v/v) fetal bovine serum, transfected with N-terminally or C-terminally tagged CALHM1, CALHM2 and CALHM3 plasmids using Lipofectamine 2000 (ThermoFisher), and placed at 37 °C for 24 h. The next day, sodium butyrate was added to each well to a final concentration of 10 mM and placed at 30 °C. Then, 24 h later, cells were imaged, collected, washed with 150 mM NaCl, and 20 mM Tris-HCl pH 8.0 buffer (TBS) and stored at –80 °C. Cell aliquots were thawed on ice, mixed with 1% digitonin-containing buffer and left to incubate on ice for 30 min. Mixtures were then centrifuged at 87,000g using a TLA100.3 rotor (Beckman Coulter) for 20 min at 4 °C. The supernatant was carefully pipetted off and injected onto a 3 ml GE Healthcare Superose 5/150 GL column, prewashed with TBS buffer plus detergent, at a flow rate of 0.3 ml/min, and analysed to determine protein solubility and retention time.

### Protein expression and purification

The following purification was revised accordingly, and carried out as previously described<sup>31–35</sup>. DNA was transformed into DH10Bac cells to produce bacmid for baculovirus production. Flasks of tsA201 suspension cells (ATCC CRL-11268, tested negative for mycoplasma contamination, and authenticated) were grown to a density of  $3.0\text{--}3.5 \times 10^6$  cells/ml at 37 °C. P2 virus (8%) (v/v) was added to each flask and incubated at 37 °C for 12 h. Sodium butyrate was added to a final concentration of 10 mM and flasks were incubated at 30 °C. Cells were collected 48 h after infection and frozen at –80 °C. For use, cells were thawed on ice and resuspended in TBS buffer supplemented with 1 mM PMSF, 0.8 μM aprotinin, 2 μg/ml leupeptin, and 2 mM pepstatin A; then, they were lysed via sonication for 15 min. The lysate was centrifuged at 7,000g for 10 min. The supernatant was ultracentrifuged at 186,000g in a Ti45 rotor (Beckman Coulter) at 4 °C for 1 h. Membrane pellets were homogenized in TBS supplemented with protease inhibitors and solubilized in 1% digitonin at 4 °C for 2 h. Membrane debris was removed by centrifugation at 186,000g in a Ti45 rotor at 4 °C for 1 h.

The solubilized protein was applied to 10 ml of TALON resin pre-equilibrated with 3 column volumes of TBS supplemented with 10 mM imidazole and 0.1% digitonin (buffer A). The column was washed with 5 column volumes of buffer A, and eluted with 3 column volumes of elution buffer containing 250 mM imidazole, pH 8.0. Fractions containing the N terminus or C terminus of CALHM2 were pooled and were concentrated directly for size-exclusion chromatography in TBS supplemented with 0.1% digitonin, pH 8.0. The initial purification using N-terminally or C-terminally GFP-tagged constructs used in cryo-EM yielded micrographs containing only hemichannels, and the three-dimensional reconstructions indicated fuzzy tails hanging outside of the detergent micelle, resulting in low-resolution reconstructions. To eradicate the flexible GFP tag, thrombin digestion was implemented

overnight at 4 °C at a ratio of 1:20 thrombin:protein. The resulting peak fractions containing GFP-cleaved CALHM2C were then combined and concentrated to 8–9 mg/ml or tested directly via negative-stain grids. GFP-cleaved CALHM2C (the C-terminally GFP-tagged construct) in the presence of EDTA or RUR was frozen for further cryo-EM studies. A high concentration of RUR destabilizes CALHM2, based on a stability test using fluorescence-detection size-exclusion chromatography (FSEC) (Extended Data Fig. 1j). Therefore, we chose a RUR concentration of 1.5 mM.

### Cryo-EM sample preparation and data acquisition

Concentrated CALHM2C protein was preincubated with EDTA or RUR for 30 min on ice. The protein mixture (2.5 μl) was then applied to glow-discharged Quantifoil carbon grids (gold, 1.2/1.3-μm size/hole space, 300 mesh), blotted for 2 s at 100% humidity using a Vitrobot Mark III, and flash-frozen in vitreous liquid ethane. Particle images were collected using the FEI Titan Krios electron microscope equipped with a nominal magnification 130,000× Gatan K2 Summit direct electron detector, recording image stacks in super-resolution counting mode at a binned pixel size of 1.026 Å. Each image was dose-fractionated in 40 frames using a total exposure time of 8 s at 0.2 s per frame. The dose rate was  $6.76 \text{ e}^- \text{Å}^{-2} \text{s}^{-1}$ . All image stacks were collected using SerialEM<sup>36</sup>, an automated acquisition program. Nominal defocus values varied from 1.0 to 2.5 μm.

### Cryo-EM data processing

Movies were motion-corrected using MotionCor2<sup>37</sup>. Gctf<sup>38</sup> was applied to non-dose-weighted micrographs to estimate defocus values. Particles were picked using Gautomatch (<http://www.mrc-lmb.cam.ac.uk/kzhang/Gautomatch/>). Templates were generated from the initial pilot results and subjected to two rounds of two-dimensional classification using RELION 2.1 and RELION 3.0-beta-2 (ref. <sup>39</sup>). In the EDTA dataset, CryoSPARC<sup>40</sup> was used to separate hemichannels and gap junctions from two-dimensional classification results. In addition, CryoSPARC was used to obtain the initial models of the hemichannel and gap junction. For the EDTA–CALHM2 data, the hemichannel particles and gap-junction particles were further cleaned up by two-dimensional classification, and the selected particles from two-dimensional classification were subjected to three-dimensional classification in RELION using maps from CryoSPARC low-pass-filtered of 60 Å as reference models. For the RUR–CALHM2 data, the EDTA–CALHM2 hemichannel map was used as a reference model during three-dimensional classification. Particles from classes that showed high-resolution features were refined without applying symmetry, and particles from classes showing obvious  $C_{11}$  symmetry were further refined using  $C_{11}$  symmetry. In the RUR–CALHM2 dataset, in addition to the symmetric class that yielded the highest-resolution structure, two well-defined non-symmetric classes were observed. Both non-symmetric classes are ellipse-shaped when viewed perpendicular to the membrane, having the S1 helices in the lifted conformation. In one of these two non-symmetric classes, prominent densities belonging to two subunits extending to the centre of the pore were observed. Notably, these two subunits are located on the opposite sides of the longer axis of the ellipse; it is possible that the ellipse shape is a result of the push between the S1 helix and NTH of these two subunits. To assess the structural heterogeneity of the first transmembrane helix S1, we analysed the particles that yielded maps with highest resolutions of RUR–CALHM2 and EDTA–CALHM2 using an approach that combined symmetry expansion and signal subtraction, in which all of the subunits were subtracted and classified without image alignment in RELION. For RUR–CALHM2, two different conformations of helix S1 appeared with 70% in the lifted conformation and 13% in the vertical conformation attached to S3, with the rest not being well-defined. No RUR densities are observed in the class with the vertical S1. For EDTA–CALHM2<sup>hem</sup>, the three-dimensional classification of a single subunit reveals that approximately 48% of the S1 helices are



in the vertical conformation and attached to S3, whereas the rest are not well-defined. We suggest that this may be caused by the high flexibility of helix S1 in the absence of RUR, and that the non-resolved densities may present the intermediate states of S1 between the lifted and vertical conformation. By contrast, three-dimensional classification of single subunits of EDTA–CALHM2<sup>gap</sup> did not yield meaningful results. The proportion of the single subunits in a gap junction structure is probably too small to perform reliable signal subtraction and subsequent three-dimensional classification.

### Model building

De novo model building of RUR–CALHM2 was carried out using Coot<sup>41</sup>, guided by bulky residues and secondary structure prediction. The architecture of CALHM2 is mainly  $\alpha$ -helices, which greatly assisted in registering assignment. The EDTA–CALHM2 models were built based on the RUR–CALHM2 model. Model building of the less-well-defined regions—including the NTH, S1 helix and part of the extracellular loops—were carried out using the maps refined without a soft solvent mask, and the maps of the single subunit obtained through symmetry expansion, signal subtraction and three-dimensional classification. The models were then subjected to real-space refinement using phenix.refine<sup>42</sup> with secondary structure restraints. The refined models were manually examined and adjusted in Coot. Although we can fit side chains in most parts of the protein, owing to intrinsic positional uncertainty we fitted only the protein backbone into the densities of the NTH, S1 helix and part of the extracellular loops based on secondary structure prediction, and our electrophysiology experiment on a mutant (CALHM2(E37R)) located in S1. The exact position of residues in these regions (residues 13–40 and 138–152) should be interpreted with caution. For validation of the refined structures, Fourier shell correlation<sup>39</sup> curves were applied to calculate the difference between the final model and electron microscopy map. The geometries of the atomic models were evaluated using MolProbity<sup>43</sup>. Figures were created using PyMOL<sup>44</sup> and UCSF Chimera<sup>45</sup>.

### Thermostability experiment

Wild-type CALHM2 and mutants were transfected into tsA201 cells as described in the ‘Construct screening’ section. After collection, wild-type CALHM2 and mutants were extracted in TBS buffer supplemented with 10 mM *n*-dodecyl  $\beta$ -D-maltoside and 2 mM cholesteryl hemisuccinate Tris salt at 4 °C for 1 h. Solubilized samples were ultracentrifuged at 87,000g for 20 min at 4 °C to remove cell debris and membranes. The resulting supernatants were heated at selected temperatures for 10 min and centrifuged at 87,000g at 4 °C for 20 min to remove any aggregates. Heated supernatants were loaded onto a 3-ml GE Healthcare Superose 5/150 GL column for high-performance liquid chromatography to measure GFP intensity (excitation 488 nm and emission 510 nm) and compared to 4-°C controls.

### Electrophysiology

Flasks of tsA201 suspension cells were grown to a density of  $1.0 \times 10^6$  cells/ml at 37 °C and infected with 1–5% (v/v) human CALHM2C P2 virus. After 12 h, 5 mM sodium butyrate was added, and the cells were left to incubate at 30 °C for an additional 24 h. Infected cells were plated to a final density of  $0.3 \times 10^6$  cells/ml in a 24-well plate on microscope cover glass (Fisher), incubated for 3 h and recorded 24–48 h after infection. Whole-cell patch-clamp recordings were collected at room temperature using a HEKA EPC-10 amplifier. Cells were held at –60 mV and data were recorded at 20 kHz and filtered at 1 kHz. Inhibitor- and activator-containing buffers were applied using a two-channel theta-glass pipette. The bath solution was prepared using 140 mM NaCl, 5.4 mM KCl, 5 mM CaCl<sub>2</sub>, 1 mM MgCl<sub>2</sub>, 10 mM HEPES and 20 mM sucrose (pH 7.4), and electrodes were filled with an internal solution of 130 mM KCl, 10 mM NaCl, 1 mM CaCl<sub>2</sub>, 10 mM HEPES, and 11 mM EGTA (pH 7.3). All data were collected using Patchmaster software (HEKA).

### Deglycosylation test

In brief, 500  $\mu$ l of pelleted cells was thawed on ice, mixed with 100  $\mu$ l of 1% digitonin detergent and left to incubate for 45 min at 4 °C. Detergent–cell mixtures were then centrifuged at 21,000g for 10 min at 4 °C. Supernatant was transferred into fresh Eppendorf tubes and centrifuged at 87,000g for 20 min at 4 °C. High-speed supernatant was then transferred into fresh Eppendorf tubes for a second time. Each sample was divided into four tubes; the first two tubes as the control and the other two tubes mixed with 0.5  $\mu$ l (250 U) of endoglycosidase H (NEB P0702S) or PNGase F (NEB P0704S), respectively. Samples were left nutating overnight at 4 °C at optimum pH. The following day, all samples were centrifuged at 186,000g for 20 min at 4 °C, mixed with 5 $\times$  sample loading buffer to a final concentration of 1 $\times$  and run on a 4–20% SDS–PAGE gel with protein standard. Unstained gel was analysed using a fluorescent gel imager to detect GFP.

### Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

### Data availability

The cryo-EM density map and coordinates of EDTA–CALHM2<sup>hemi</sup>, EDTA–CALHM2<sup>gap</sup>, and RUR–CALHM2 have been deposited in the Electron Microscopy Data Bank (EMDB) under accession numbers EMDB-20788, EMDB-20790 and EMDB-20789, respectively, and in the Research Collaboratory for Structural Bioinformatics Protein Data Bank under accession codes 6UIV, 6UIX and 6UIW, respectively. The single subunit map(s) obtained from signal subtraction and associated mask have been deposited under the corresponding EMDB accession number.

- Goehring, A. et al. Screening and large-scale expression of membrane proteins in mammalian cells for structural studies. *Nat. Protocols* **9**, 2574–2585 (2014).
- Winkler, P. A., Huang, Y., Sun, W., Du, J. & Lü, W. Electron cryo-microscopy structure of a human TRPM4 channel. *Nature* **552**, 200–204 (2017).
- Huang, Y., Roth, B., Lü, W. & Du, J. Ligand recognition and gating mechanism through three ligand-binding sites of human TRPM2 channel. *eLife* **8**, e50175 (2019).
- Fan, C., Choi, W., Sun, W., Du, J. & Lü, W. Structure of the human lipid-gated cation channel TRPC3. *eLife* **7**, e36852 (2018).
- Huang, Y., Winkler, P. A., Sun, W., Lü, W. & Du, J. Architecture of the TRPM2 channel and its activation mechanism by ADP-ribose and calcium. *Nature* **562**, 145–149 (2018).
- Haley, E. et al. Expression and purification of the human lipid-sensitive cation channel TRPC3 for structural determination by single-particle cryo-electron microscopy. *J. Vis. Exp.* **143**, e58754 (2018).
- Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).
- Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
- Zhang, K. Gctf: real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
- Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
- Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
- Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
- Afonine, P. V. et al. Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D* **68**, 352–367 (2012).
- Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
- Delano, W. The PyMOL Molecular Graphics System. <https://pymol.org/> (2002).
- Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
- Heymann, J. B. & Belnap, D. M. Bsoft: image processing and molecular modeling for electron microscopy. *J. Struct. Biol.* **157**, 3–18 (2007).
- Drozdetskiy, A., Cole, C., Procter, J. & Barton, G. J. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* **43**, W389–W394 (2015).

**Acknowledgements** We thank G. Zhao and X. Meng for the support with data collection at the David Van Andel Advanced Cryo-Electron Microscopy Suite; the HPC team of VARI for computational support; D. Nadziejka for technical editing and E. Haley for proofreading. J.D. is supported by a McKnight Scholar Award, a Klingenstein-Simon Scholar Award and the National Institutes of Health (NIH) (grant R01NS111031).

# Article

**Author contributions** W.L. and J.D. initiated the project. W.C. and N.C. purified CALHM2, and prepared and screened cryo-EM samples. N.C., W.C. and W.S. performed functional studies. W.C., J.D. and W.L. performed cryo-EM data collection and processing. W.L. and J.D. performed data analysis and wrote the manuscript. All of the authors contributed to manuscript preparation.

**Competing interests** The authors declare no competing interests.

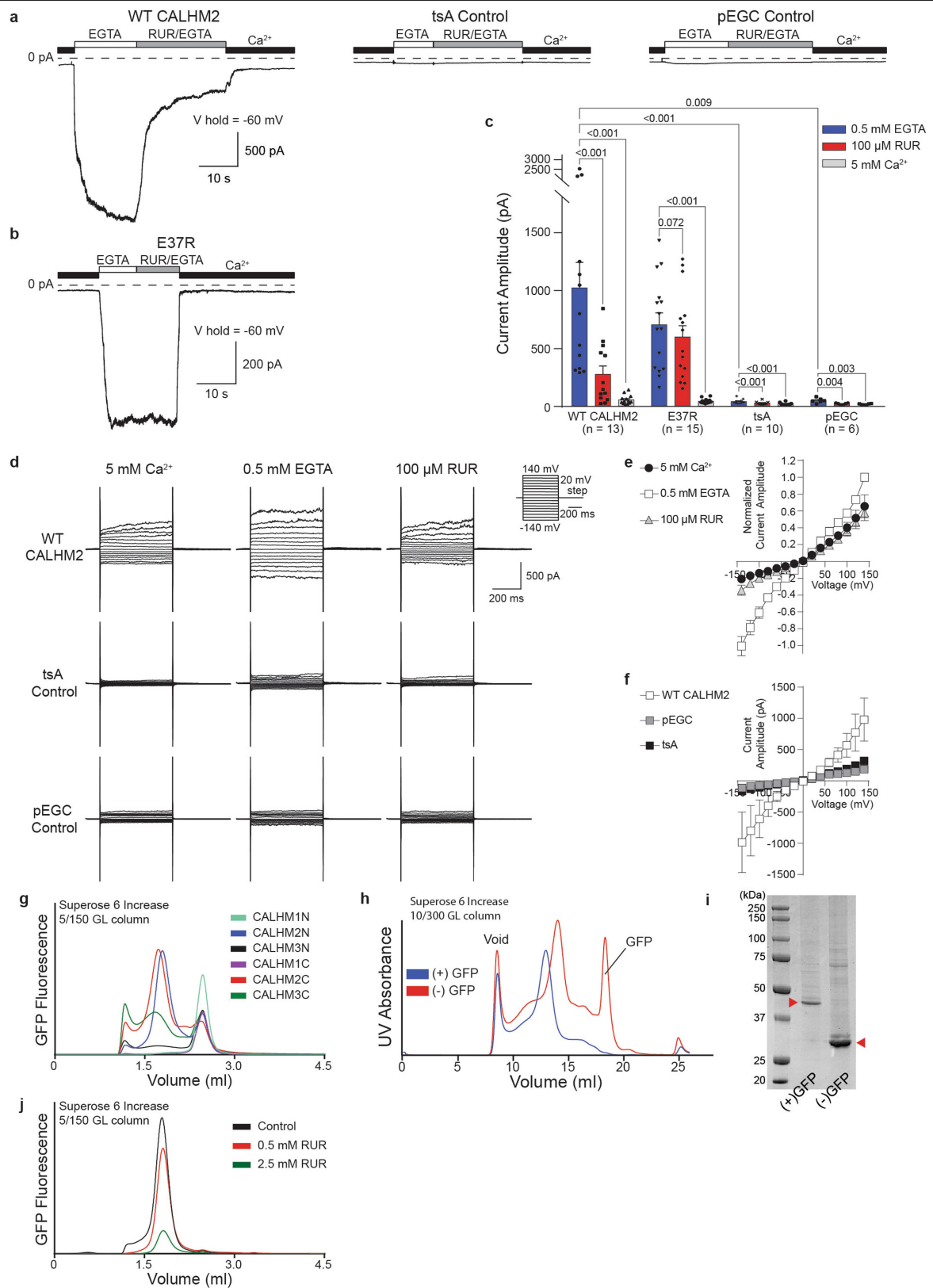
## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1781-3>.

**Correspondence and requests for materials** should be addressed to J.D. or W.L.

**Peer review information** *Nature* thanks Kenton Swartz and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

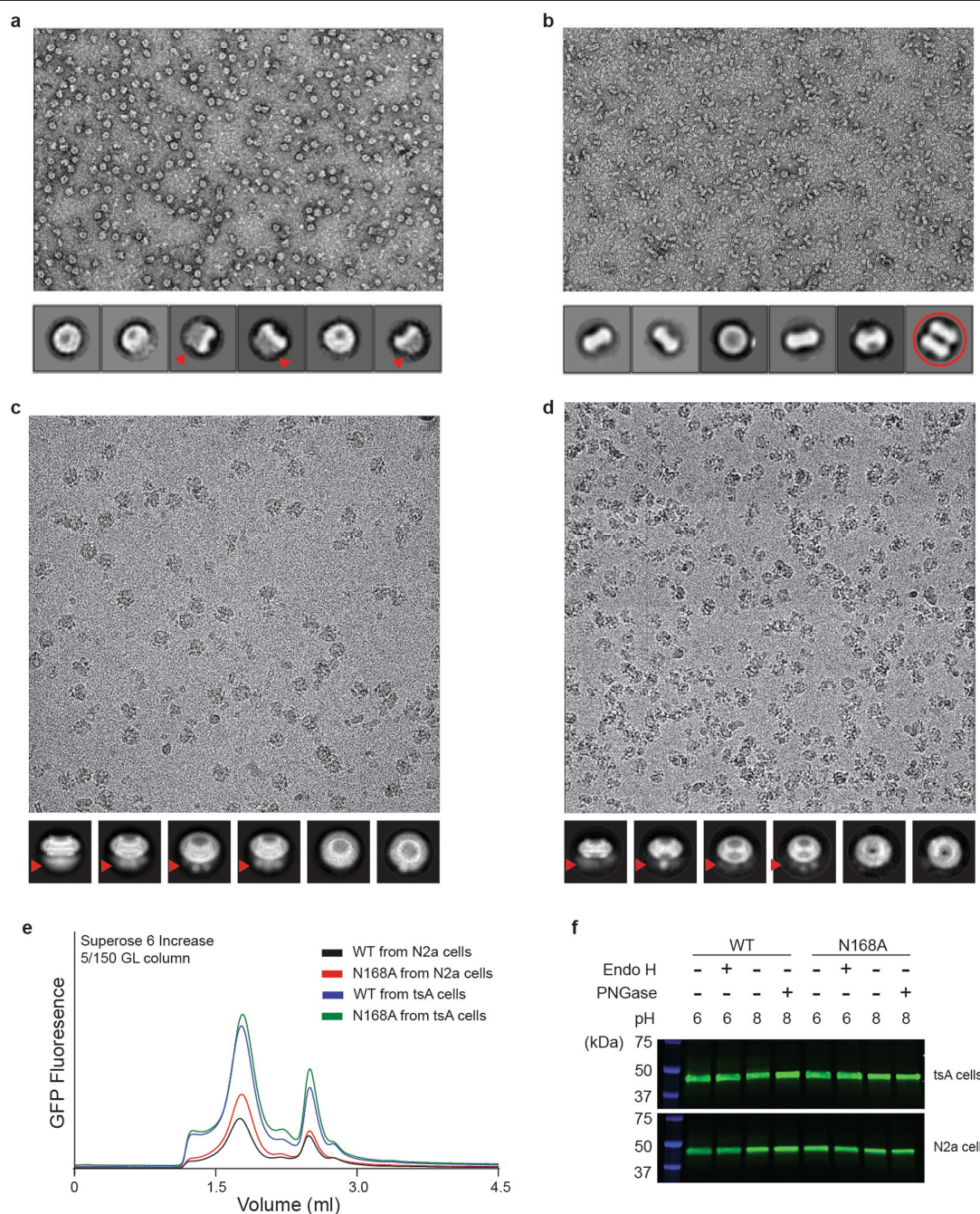


**Extended Data Fig. 1** | See next page for caption.

# Article

**Extended Data Fig. 1 | Electrophysiology experiments, construct screening and purification.** **a, b**, Representative current traces recorded in whole-cell mode at  $-60$  mV for cells expressing wild-type CALHM2, tsA control cells, tsA cells transfected with empty pEGC vector (**a**) or tsA cells expressing CALHM2(E37R) (**b**). Cells were switched from bath buffer that contained 5 mM  $\text{Ca}^{2+}$  to one that contained 0.5 mM EGTA (0 mM  $\text{Ca}^{2+}$ ) to induce the current. The current was inhibited using a buffer that contained 100  $\mu\text{M}$  RUR and 0.5 mM EGTA.  $n = 13, 15, 10$  or 6 biologically independent experiments were performed for wild-type CALHM2, CALHM2(E37R), tsA control or pEGC control, respectively. **c**, Quantification of current amplitude in 0.5 mM EGTA, 100  $\mu\text{M}$  RUR and 0.5 mM EGTA, and 5 mM  $\text{Ca}^{2+}$  conditions for CALHM2-expressing and control cells from **a, b**. Two-tailed paired and unpaired  $t$ -tests were applied to calculate  $P$  values (using GraphPad Prism 7), from within and outside of each cell type, respectively. Each dot indicates the value of one single independent experiment. RUR inhibition was nearly abolished in the CALHM2(E37R) mutant. **d**, Representative current–voltage relationships obtained by applying 500-ms voltage pulses ranging from 140 mV to  $-140$  mV from a holding potential of 0 mV (20-mV steps) to cells expressing wild-type CALHM2 (top), tsA control cells (middle) and tsA cells transfected with empty pEGC vector (bottom). Currents were recording in the presence of 5 mM  $\text{Ca}^{2+}$ , 0.5 mM EGTA (0 mM  $\text{Ca}^{2+}$ ) or 100  $\mu\text{M}$  RUR and 0.5 mM EGTA.  $n = 7, 6$  or 4 biologically

independent experiments were performed for wild-type CALHM2, tsA control or pEGC control, respectively. **e**, The data of wild-type CALHM2 in **d** were normalized to the amplitude of the current recorded in the presence of EGTA at 140 mV and calculated as mean  $\pm$  s.e.m. from 7 cells. **f**, Averaged current amplitude of cells expressing wild-type CALHM2 ( $n = 7$ ), cells transfected with empty pEGC vector ( $n = 4$ ) and tsA control cells ( $n = 6$ ) in the presence of 0.5 mM EGTA, plotted as mean  $\pm$  s.e.m. **g**, FSEC profiles showing that human CALHM2 N-terminally tagged with GFP (CALHM2N) and human CALHM2 C-terminally tagged with GFP (CALHM2C) have the best biochemical properties among human CALHM1, CALHM2 and CALHM3 when solubilized in digitonin. This experiment was repeated three times yielding similar results. **h**, Size-exclusion chromatography profiles of CALHM2C in digitonin before (blue) and after (red) GFP cleavage. After cleavage, the main peak shifted towards a smaller molecular weight. Purification of CALHM2C was repeated multiple times ( $>10$ ), yielding similar results in each case. **i**, SDS gel of purified CALHM2C (indicated by red arrowheads) before (left band) and after (right band) GFP cleavage. **j**, Stability test of purified CALHM2C in the presence of RUR at two concentrations using FSEC, showing that a high concentration of RUR decreases CALHM2 stability. This experiment was repeated five times, each yielding similar results.

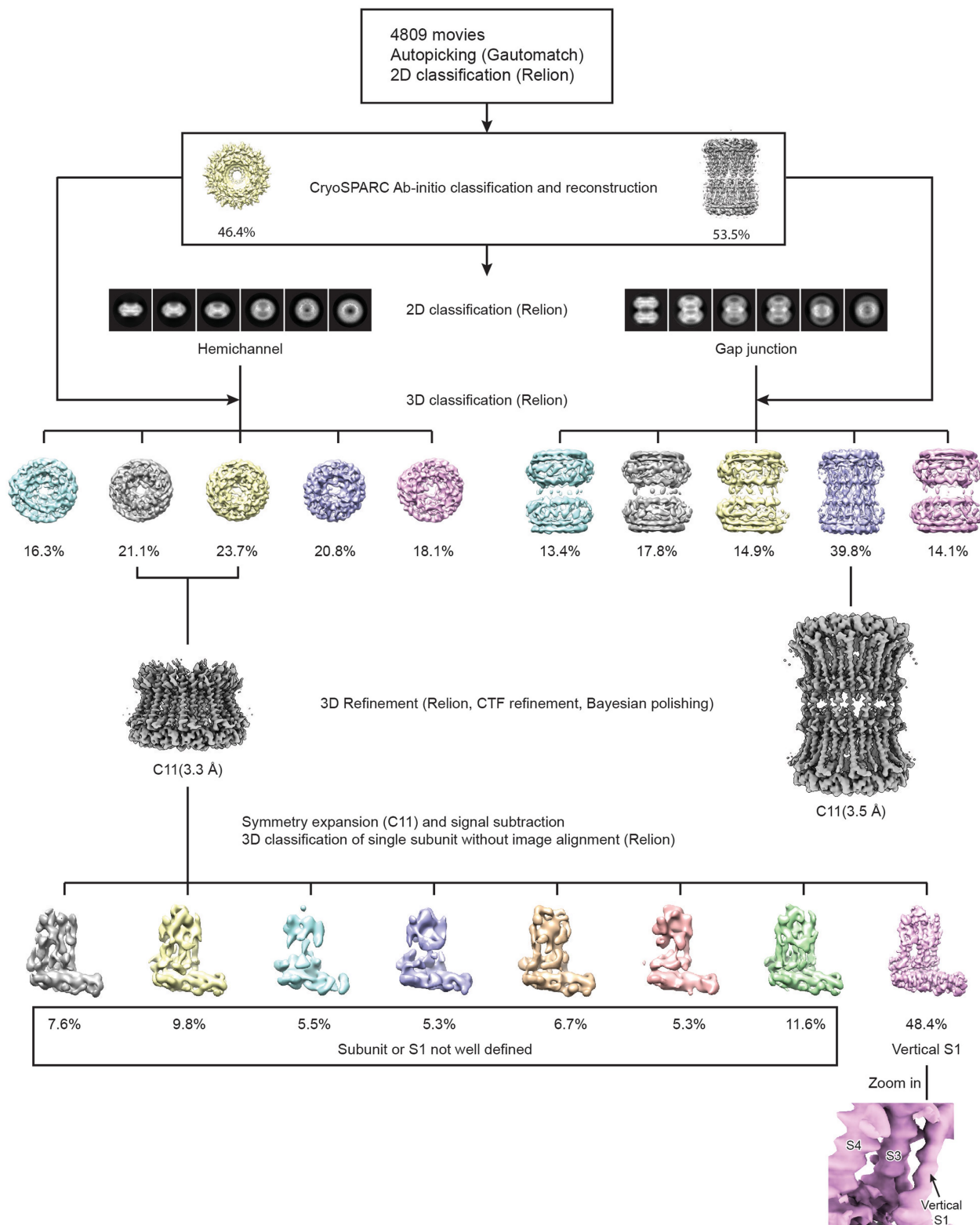


# **Extended Data Fig. 2 | Factors that affect the formation of a gap junction.**

**a**, A representative micrograph of purified CALHM2N by negative-stain electron microscopy; selected two-dimensional classes are shown below. Fuzzy areas (indicated by red arrowheads) are caused by flexible GFP tags; this is also implied in **c** and **d**. Only hemichannels, and not gap junctions, appear in the micrograph. **b**, A representative micrograph of purified CALHM2C (from which the GFP tag has been cleaved) by negative-stain electron microscopy. The fuzzy areas in **a**, **c**, **d** disappeared in the two-dimensional classes, which confirms that they are indeed due to the GFP tag. One of the two-dimensional class averages represents a gap junction (red circle). **c**, A representative micrograph of purified CALHM2N by cryo-EM in the presence of 1 mM EDTA, collected using the Titan Krios; selected two-dimensional classes are shown

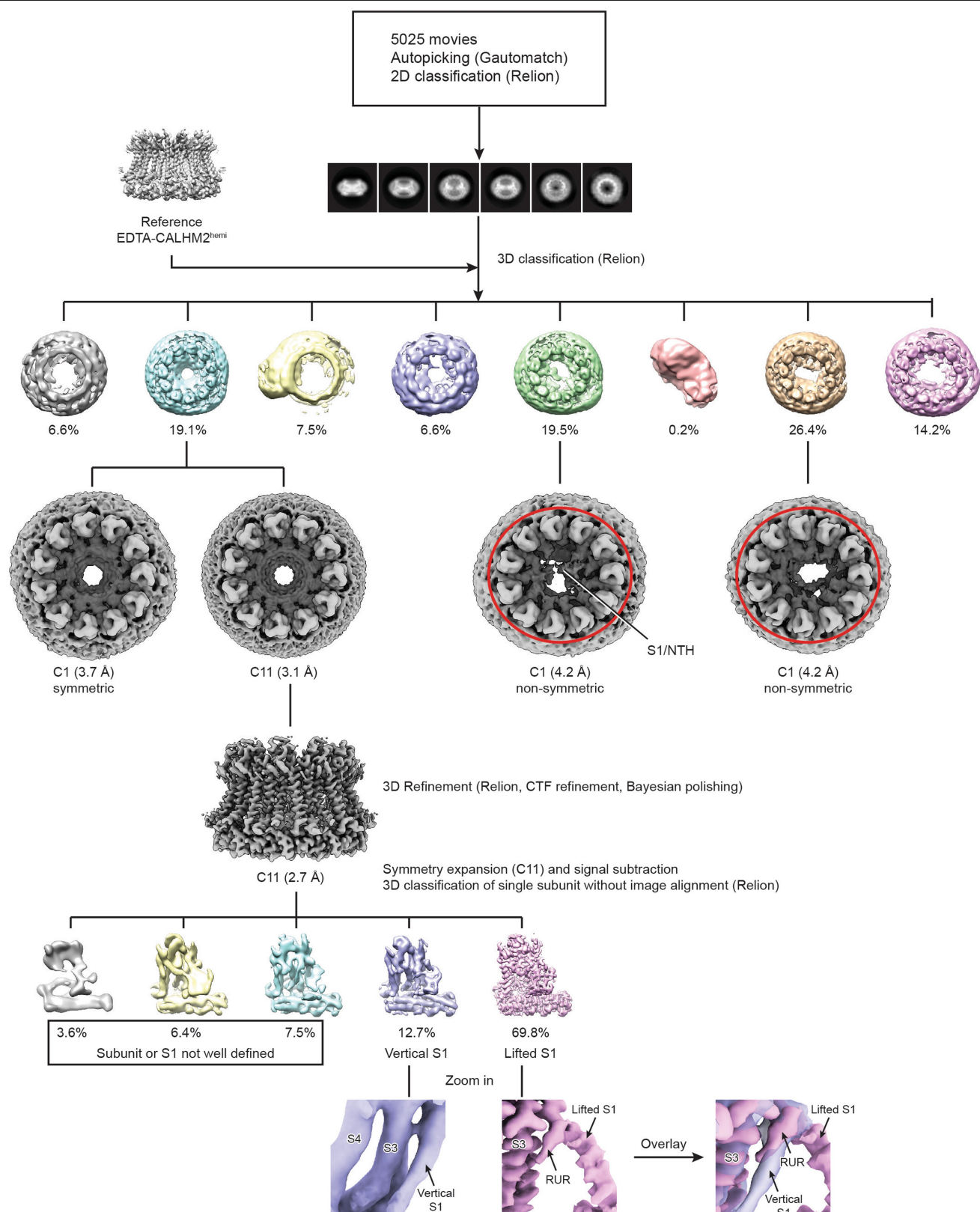
below. Only hemichannels, and not gap junctions, are observed. **d**, A representative micrograph of purified CALHM2C by cryo-EM using the Talos Arctica; selected two-dimensional classes are shown below. Only hemichannels, and not gap junctions, are seen. **e**, FSEC profile of wild-type CALHM2 and the CALHM2(N168A) mutant, expressed in tsA 201 and N2a cells. This experiment was repeated three times, each yielding similar results. **f**, The wild-type CALHM2 ran at the same height in SDS gel before and after treatment using 250 U of Endo H and 250 U PNGase at optimum pH. Moreover, the CALHM2(N168A) mutant ran at the same height as the wild-type CALHM2. These data suggest CALHM2 is not glycosylated. This experiment was repeated multiple times (Endo H,  $n = 6$ ; PNGase,  $n = 3$ ), all yielding non-glycosylation phenotypes.





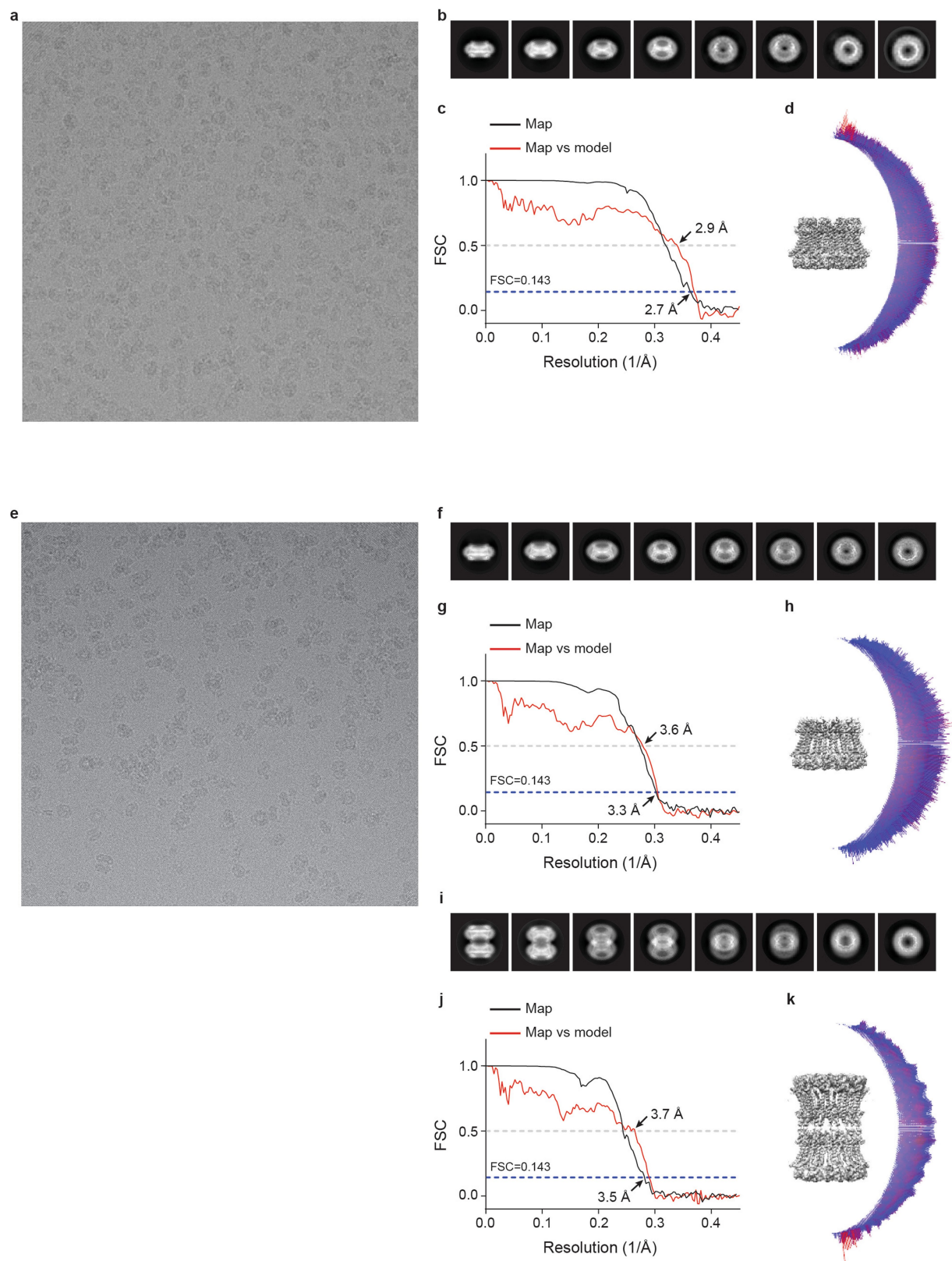
**Extended Data Fig. 3 | The workflow of cryo-EM data processing for EDTA-CALHM2.** A total of 4,809 movies was collected using a Titan Krios equipped with K2. Particles were autopicked using Gautomatch, and visually examined in RELION to eradicate false-positive selections. After manual clean-up, particles were subjected to two rounds of two-dimensional classification in RELION. Ab initio reconstruction with two classes was performed in CryoSPARC to separate hemichannel and gap-junction particles, and to generate initial models of the hemichannel and gap junction. Hemichannel and gap-junction

particles were further cleaned up using two-dimensional and three-dimensional classification in RELION. Particles from three-dimensional classes that showed high-resolution features and obvious  $C_{11}$  symmetry were combined and refined in RELION. To assess the structural heterogeneity of the helix S1, an approach that combined symmetry expansion and signal subtraction was carried out, in which all of the subunits were subtracted and classified without image alignment in RELION.



**Extended Data Fig. 4 | The workflow of cryo-EM data processing for RUR-CALHM2.** A total of 5,025 movies was collected using a Titan Krios equipped with K2. Particles were autopicked using Gautomatch, and visually examined in RELION to eradicate false-positive selections. After manual clean-up, particles were subjected to two rounds of two-dimensional classification in RELION. Three-dimensional classification using the map of EDTA-CALHM2<sup>hemi</sup> as an initial model yielded three classes with high-resolution features. Particles from the three classes were refined without applying symmetry, and particles from

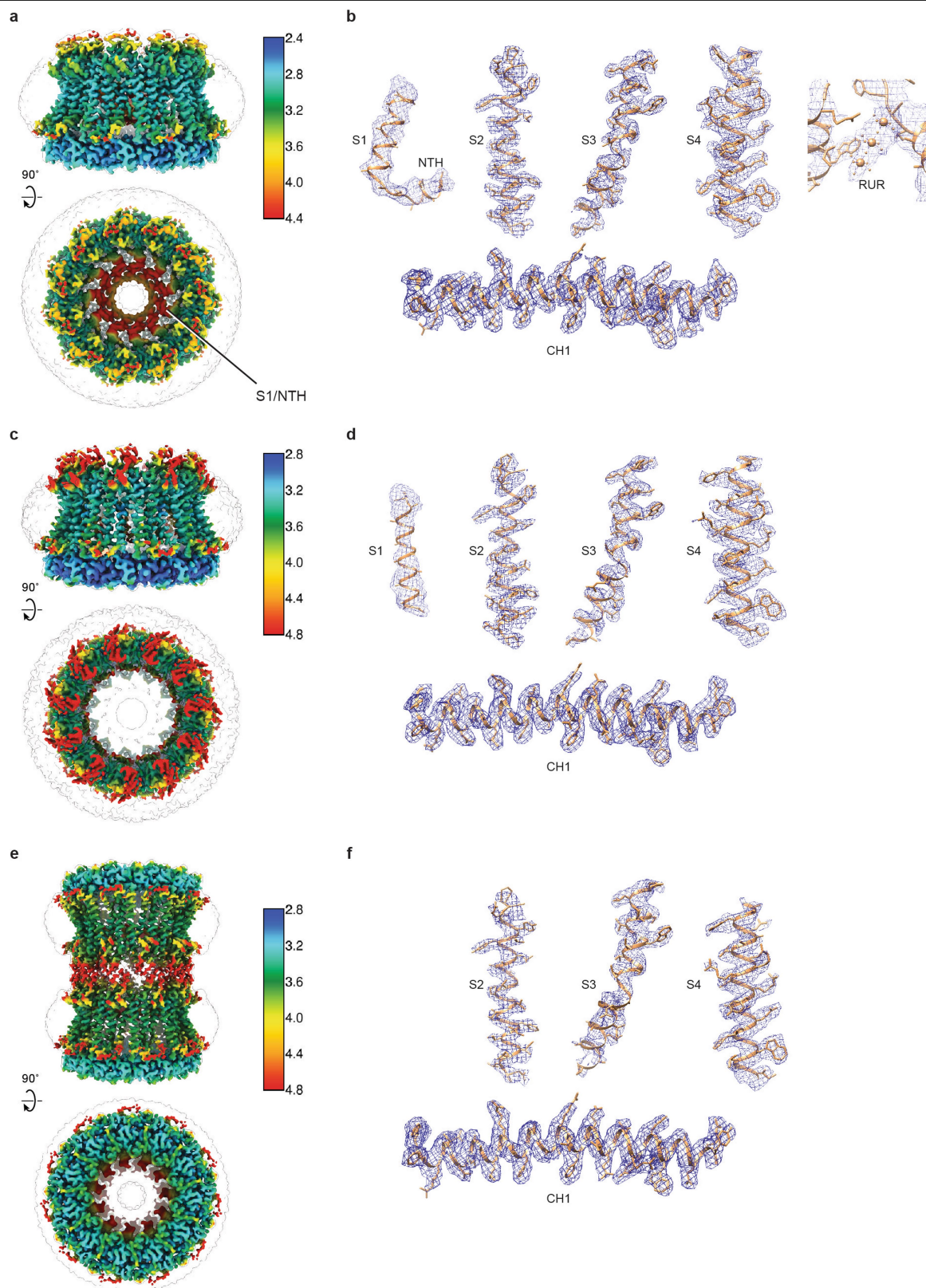
the class that showed obvious C<sub>11</sub> symmetry were further refined using C<sub>11</sub> symmetry. The two non-symmetric classes are highlighted by the red ellipses. The densities of the S1 helix and NTH that extend to the pore centre in one of the non-symmetric class are labelled. To assess the structural heterogeneity of helix S1, an approach that combined symmetry expansion and signal subtraction was carried out, in which all of the subunits were subtracted and classified without image alignment in RELION.



**Extended Data Fig. 5 | Cryo-EM analysis of human CALHM2.** **a, e,** Representative electron micrograph of RUR-CALHM2 (**a**, out of 5,025 micrographs) and EDTA-CALHM2 (**e**, out of 4,809 micrographs). **b, f, i,** Selected two-dimensional class averages of the electron micrographs of RUR-CALHM2 (**b**), EDTA-CALHM2<sup>hemi</sup> (**f**) and EDTA-CALHM2<sup>gap</sup> (**i**). **c, g, j,** The gold-standard Fourier shell correlation (FSC) curves for the electron microscopy maps of RUR-CALHM2

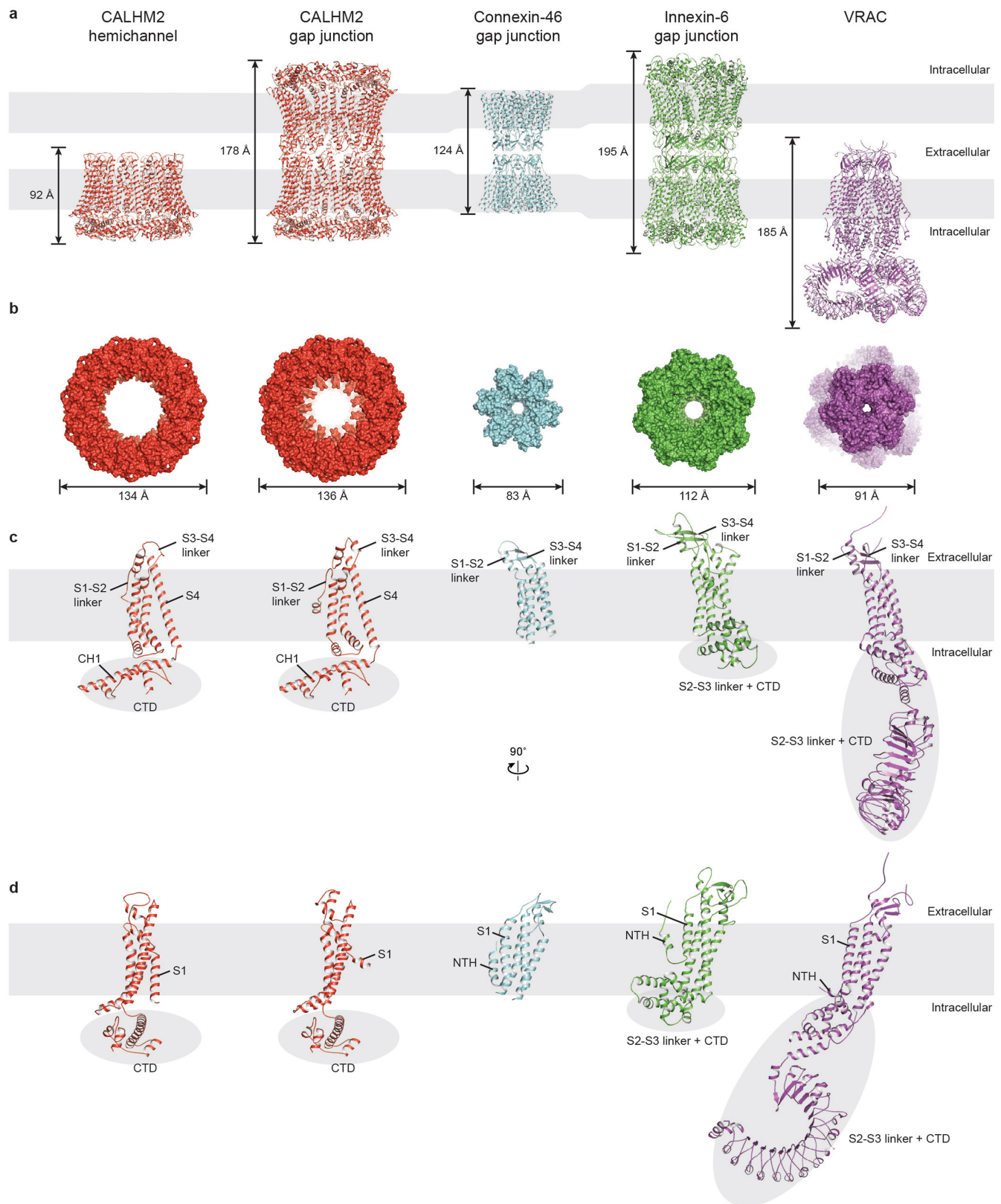
(**c**), EDTA-CALHM2<sup>hemi</sup> (**g**) and EDTA-CALHM2<sup>gap</sup> (**j**) are shown in black, and the Fourier shell correlation curves between the atomic model and the final electron microscopy map are shown in red. **d, h, k,** The angular distribution of particles used for the refinement of RUR-CALHM2 (**d**), EDTA-CALHM2<sup>hemi</sup> (**h**) and EDTA-CALHM2<sup>gap</sup> (**k**).





**Extended Data Fig. 6 | Representative densities of the reconstructions of RUR-CALHM2, EDTA-CALHM2<sup>hemi</sup> and EDTA-CALHM2<sup>gap</sup>.** **a, c, e**, Local-resolution estimation of the structure of RUR-CALHM2 (**a**), EDTA-CALHM2<sup>hemi</sup> (**c**) and EDTA-CALHM2<sup>gap</sup> (**e**), calculated using Bsoft<sup>46</sup>. **b, d, f**, Representative densities of RUR-CALHM2 (**b**), EDTA-CALHM2<sup>hemi</sup> (**d**) and EDTA-CALHM2<sup>gap</sup> (**f**). The putative RUR-binding site density is shown in the panel on the right in **b**.

(c) and EDTA-CALHM2<sup>gap</sup> (**e**), calculated using Bsoft<sup>46</sup>. **b, d, f**, Representative densities of RUR-CALHM2 (**b**), EDTA-CALHM2<sup>hemi</sup> (**d**) and EDTA-CALHM2<sup>gap</sup> (**f**). The putative RUR-binding site density is shown in the panel on the right in **b**.



**Extended Data Fig. 7 | Comparison of EDTA-CALHM2<sup>hemi</sup> and EDTA-CALHM2<sup>gap</sup> with the connexin-46 gap junction, innexin-6 gap junction and a VRAC. **a, b**, Overall structure comparison viewed parallel to the membrane (**a**, cartoon representation) and viewed from the intracellular side (**b**, surface representation), showing notable differences in symmetry, size and shape. The VRAC in **b** is viewed from the extracellular side. The size of the VRAC in **b** represents the largest diameter of the transmembrane domain. **c, d**, Single-**

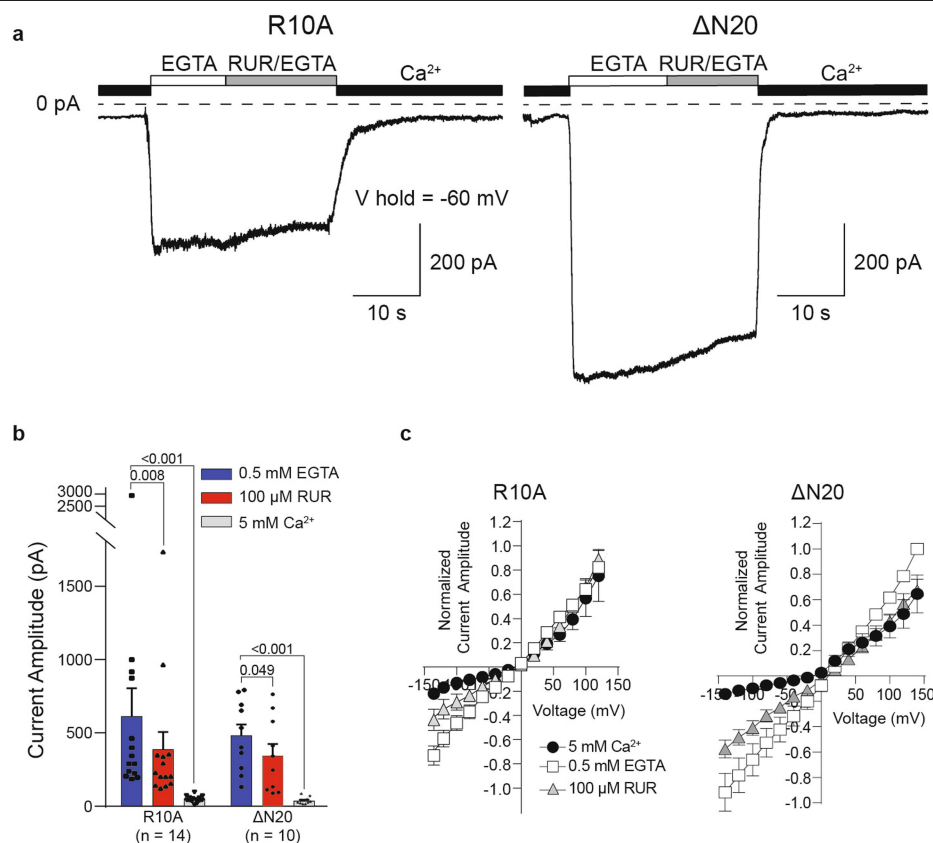
subunit comparison in two different views viewed parallel to the membrane. The intracellular domains are highlighted using grey ellipses, and the grey rectangles represent cell membranes. The buried surface area in each pair of protomers in the CALHM2 gap junction is 378 Å<sup>2</sup> (4,161 Å<sup>2</sup> for an undecamer), which is substantially smaller than the equivalent buried surface area in connexin (94 Å<sup>2</sup>; or 5,654 Å<sup>2</sup> for a hexamer) and innexin (1,550 Å<sup>2</sup>; or 12,397 Å<sup>2</sup> for an octamer).





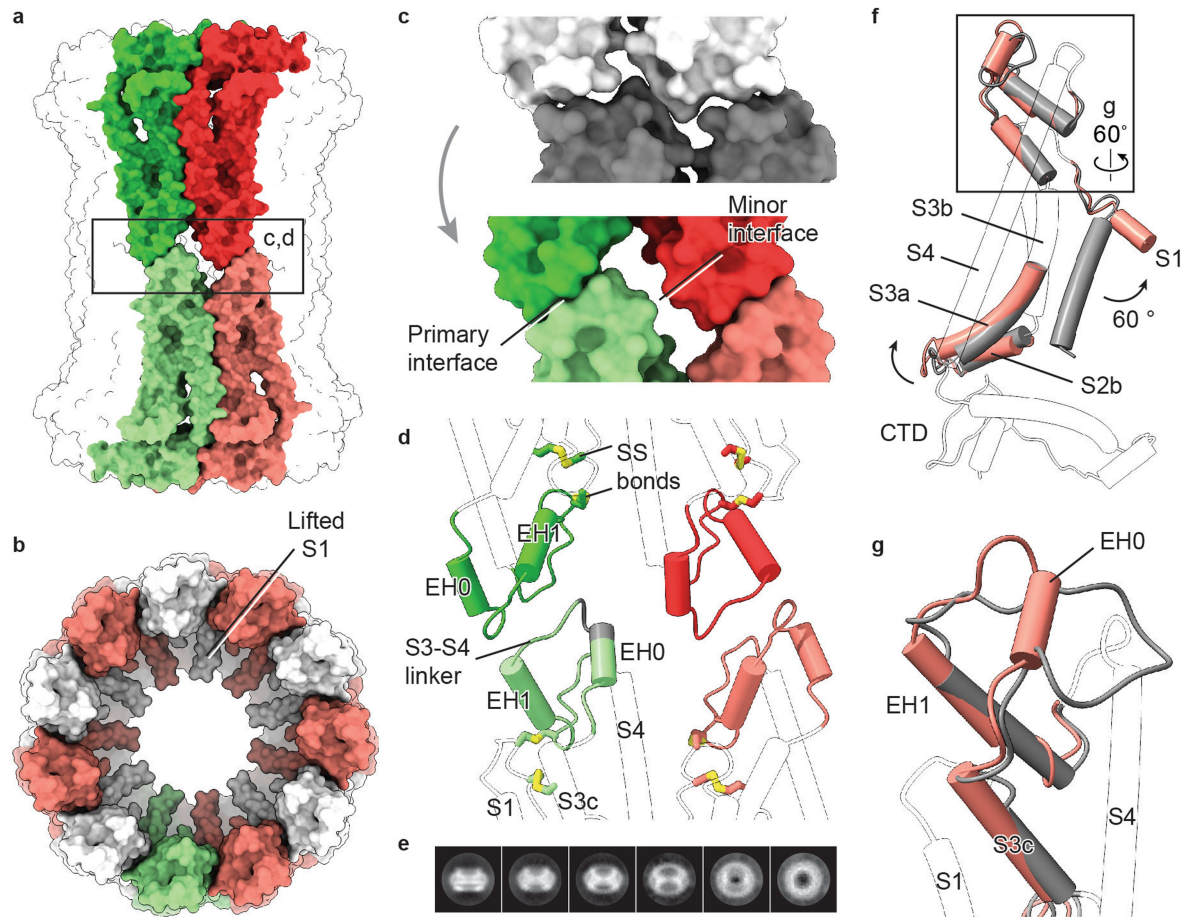
**Extended Data Fig. 8 | Secondary structure arrangement and domain organization of human CALHM2, and sequence alignment of the human CALHM family. a,** The secondary structure prediction of human CALHM2, and sequence alignment of CALHM family members CALHM1, CALHM2, CALHM3, CALHM4, CALHM5 and CALHM6. Secondary structure prediction was performed using the JPred online server<sup>47</sup>. Sequences were aligned using the

Clustal Omega program and coloured using BLOSUM62 by conservation. Residues involved in RUR binding are marked with black filled circles. The extracellular helix (EH)0 in the S3–S4 linker is formed upon the docking of two hemichannels. P86 in human CALHM1 is boxed in red. **b,** Domain organization of the human CALHM2 protomer.



**Extended Data Fig. 9 | The role of NTH in inhibition by RUR.** **a**, Representative current traces recorded in whole-cell mode at  $-60$  mV in cells expressing CALHM2(R10A) or CALHM2( $\Delta N20$ ). Human CALHM2 has three positively charged and two negatively charged residues in the NTH, resulting in one net positive charge. Out of these charged residues, R10 is the only residue that is conserved across CALHM1, CALHM2 and CALHM3. Cells were switched from bath buffer that contained 5 mM  $Ca^{2+}$  to one that contained 0.5 mM EGTA (0 mM  $Ca^{2+}$ ) to induce current. Current was inhibited using a buffer that contained 100  $\mu M$  RUR and 0.5 mM EGTA.  $n = 14$  or 10 biologically independent experiments were performed for CALHM2(R10A) or CALHM2( $\Delta N20$ ), respectively. **b**, Quantification of current amplitude in 0.5 mM EGTA, 100  $\mu M$  RUR and 0.5 mM EGTA, and 5 mM  $Ca^{2+}$  conditions for cells expressing

CALHM2(R10A) or CALHM2( $\Delta N20$ ), from **a**. Two-tailed paired  $t$ -tests were applied to calculate  $P$  values for comparisons using GraphPad Prism 7. Data are mean  $\pm$  s.e.m. Each dot indicates the value of one single independent experiment. **c**, Current-voltage relationships were obtained by applying 500-ms voltage pulses that ranged from 140 to  $-140$  mV from a holding potential of 0 mV (20-mV steps) to cells that express CALHM2(R10A) or CALHM2( $\Delta N20$ ). Currents were recorded in the presence of 5 mM  $Ca^{2+}$ , 0.5 mM EGTA and 100  $\mu M$  RUR. Data were normalized to the amplitude of the current recorded in the presence of EGTA at 140 mV, and calculated as mean  $\pm$  s.e.m.  $n = 7$  biologically independent experiments were performed each for CALHM2(R10A) and CALHM2( $\Delta N20$ ).



**Extended Data Fig. 10 | The CALHM2 gap junction.** **a**, Surface representation of a gap junction viewed parallel to the membrane. Two paired subunits are highlighted. **b**, Surface representation of a hemichannel in the gap junction, viewed from extracellular side. **c**, Interface remodelling when docking two hemichannels (shown in grey; top) into a gap junction (shown in colour; bottom). **d**, Cartoon representation of the interface between two hemichannels. The two disulfide bonds are shown. The grey segment of S3–S4 linker represents deleted residues in a mutant (CALHM2( $\Delta$ 143–146)). Only parts involved in the docking of hemichannels and disulfide bonds are highlighted in colour. **e**, Selected two-dimensional class averages of CALHM2( $\Delta$ 143–146). This mutant yielded only hemichannels, and not gap junctions. **f**, Superimposition of single subunits of EDTA–CALHM2<sup>hemi</sup> (grey) and EDTA–CALHM2<sup>gap</sup> (pink) using the CTD. Only parts with conformational changes are highlighted in colour. To understand the conformational changes upon docking, we compared single subunits of a hemichannel and a gap junction. In the hemichannel, the loop connecting segment S3c and EH1 in the

S3–S4 linker is flat and lacks extensive contact with the rest of the protein, giving rise to a flexible area that is probably required for the initiation of docking. Indeed, the S3–S4 linker is defined better in RUR–CALHM2 than in EDTA–CALHM2<sup>hemi</sup>, by forming interactions with the adjacent subunit. We suggest that the restricted S3–S4 linker in the RUR–CALHM2 hinders the docking of the hemichannels. **g**, Enlargement of the box in **f**, showing the remodelling of the S3–S4 linker from EDTA–CALHM2<sup>hemi</sup> (grey) to EDTA–CALHM2<sup>gap</sup> (pink). Upon docking, the S3c–EH1 loop remodels into two short loops and a short  $\alpha$ -helix (EH0); the EH0–EH1 loop forms the primary interface and EH0 forms the minor interface in **d**. This motion accompanies an elevation of the S3–S4 linker and segment S3c that leads to an outward flexing of S3a, which breaks the loose interface between S3a and helix S1 in **f**. As a consequence, the S1 helix is detached from S3 and moves into a lifted conformation. The conformational changes of TMD upon docking in EDTA–CALHM2 are notably consistent with those induced by RUR. Moreover, the two docked hemichannels in the gap junction have similar conformations of their S1 helices.

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☒ ☐ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection

SerialEM 3.7, Patchmaster 2x90.5

Data analysis

Gctf-1.06, Gautamatch-0.56, Relion-3.0, CryoSparc-v2, coot-0.8.9.2, pymol-2.3.2, Motioncor2-1.1.0, phenix.real\_space\_refine\_dev\_3500, phenix.molprobity\_dev\_3500, UCSF chimera\_1.13.1, UCSF chimera\_0.91, GraphPad Prism 7

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The cryo-EM density map and coordinates of EDTA-CALHM2hemi, EDTA-CALHM2gap, and RUR-CALHM2 have been deposited in the Electron Microscopy Data Bank (EMDB) under accession numbers EMDB-20788, EMDB-20790 and EMDB-20789 and in the Research Collaboratory for Structural Bioinformatics Protein Data Bank under accession codes 6UIV, 6UIX and 6UIW. The single subunit map(s) obtained from signal subtraction and associated mask have been deposited under the corresponding EMDB accession number.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences    ☐ Behavioural & social sciences    ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	All the electrophysiology experiments were repeated at least four times using different cells. The sample size was determined based on the consistence of the recordings.
Data exclusions	No data was excluded from the analysis
Replication	We have done each group of experiment with several batches of cells, different infections and with multiple independent researchers, to ensure reproducibility within the lab.
Randomization	For electrophysiology experiments, cells with GFP fluorescence (proteins were GFP-tagged) were randomly selected.
Blinding	The investigators were blinded to group allocation during data collection and analysis

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	Sf9 cells and tsA201 cells were purchased from ATCC
Authentication	Sf9 cells and tsA201 cells were authenticated
Mycoplasma contamination	Sf9 cells and tsA201 cells were tested negative form Mycoplasma contamination
Commonly misidentified lines (See <a href="#">ICLAC</a> register)	No commonly misidentified lines were used



# Author Correction: Maternal vitamin C regulates reprogramming of DNA methylation and germline development

---

<https://doi.org/10.1038/s41586-019-1699-9>

---

Correction to: *Nature* <https://doi.org/10.1038/s41586-019-1536-1>

---

Published online 04 September 2019

---

Stephanie P. DiTroia, Michelle Percharde, Marie-Justine Guerquin, Estelle Wall, Evelyne Collignon, Kevin T. Ebata, Kathryn Mesh, Swetha Mahesula, Michalis Agathocleous, Diana J. Laird, Gabriel Livera & Miguel Ramalho-Santos

---

In this Letter, draft versions of the five Supplementary Data files were inadvertently uploaded. These Supplementary Data files have all now been replaced online.

# Author Correction: Targeting cardiac fibrosis with engineered T cells

<https://doi.org/10.1038/s41586-019-1761-7>

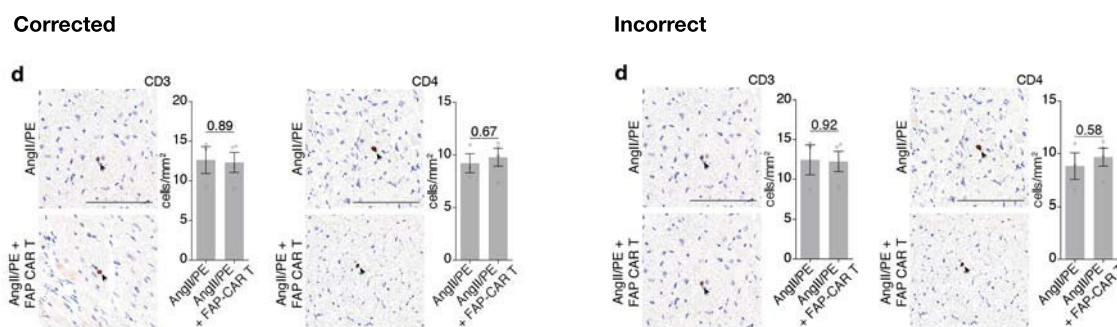
Correction to: *Nature* <https://doi.org/10.1038/s41586-019-1546-z>

Published online 11 September 2019

Haig Aghajanian, Toru Kimura, Joel G. Rurik, Aidan S. Hancock, Michael S. Leibowitz, Li Li, John Scholler, James Monslow, Albert Lo, Wei Han, Tao Wang, Kenneth Bedi, Michael P. Morley, Ricardo A. Linares Saldana, Nikhita A. Bolar, Kendra McDaid, Charles-Antoine Assenmacher, Cheryl L. Smith, Dagmar Wirth, Carl H. June, Kenneth B. Margulies, Rajan Jain, Ellen Puré, Steven M. Albelda & Jonathan A. Epstein

In this Letter, the bottom subpanel in the top left panel of Extended Data Fig. 8d, which shows CD3 staining of left ventricle tissue after AngII/PE + FAP CAR T treatment, was inadvertently duplicated from the top image of AngII/PE staining. The figure has been revised and original images re-quantified. Figure 1 of this Amendment shows the incorrect and the corrected top left panel of Extended Data Fig. 8d, for transparency to readers. Extended Data Fig. 8 and its Source Data have been corrected online. (The Supplementary Information to this Amendment contains the incorrect Source Data, for transparency to readers.)

**Supplementary Information** is available in the online version of this Amendment.



**Fig. 1** | This figure shows the corrected subpanel of the top left panel of Extended Data Fig. 8d, including the incorrect, published subpanel, for comparison.

# Publisher Correction: Harnessing innate immunity in cancer therapy

---

<https://doi.org/10.1038/s41586-019-1758-2>

---

Correction to: Nature <https://doi.org/10.1038/s41586-019-1593-5>

---

Published online 2 October 2019

---

Olivier Demaria, Stéphanie Cornen, Marc Daëron, Yannis Morel,  
Ruslan Medzhitov & Eric Vivier

---

In this Review, owing to an error during the production process, an incorrect version of Supplementary Table 1 was published. This has now been updated, and for transparency to readers the original Supplementary Table 1 is provided as Supplementary Information to this Amendment. Additionally, there were errors in the citations in the paragraph beginning ‘The interplay between innate and adaptive immunity...’. These have been amended and references 11–13 have been renumbered in the reference list accordingly. The original reference 11 (Scheper et al., Low and variable tumor reactivity of the intratumoral TCR repertoire in human cancers. *Nat. Med.* 25, 89–94 (2019)) is no longer cited in the Review, and reference 13 (Croizat et al., 2010) has been added and is cited in the above paragraph. The original Review has been corrected online.

**Supplementary Information** is available in the online version of this Amendment.



ARTHUR MEYERSON/GETTY

Researchers moving to a new country can benefit from support from local colleagues.

## SAY HELLO TO YOUR NEW BENCHMARK

How to welcome an international colleague. By Lara Pivodic

**I**n my work as a health-sciences researcher, I've moved to several countries, including Belgium, the Netherlands and the United Kingdom. I've learnt that for every researcher who moves abroad, there are several more who will welcome that individual and support their transition. In today's global research environment, which values mobility and often expects it from researchers, these local teams have a responsibility to help colleagues from abroad have a good start.

During my own international moves, I have experienced at first hand the difference that a supportive environment can make, alongside my own efforts to adapt to the new environments. Here are seven hospitable things that research teams can do to help their colleagues.

**Welcome the new team member.** First impressions count on both sides and can set the tone for days and weeks to come. It is crucial to be aware that someone is joining the group, to know the date of their first workday and to say 'welcome' in some way.

On the first day of my secondment to the United Kingdom as a PhD student, I had a meeting with my academic supervisors, one with an administrator for official intake and one with my 'buddy' – a fellow PhD student who helped me to take my first steps in the new place and introduced me to the team.

**Value input.** This step applies even if they are staying for only a short period. Invite them to all activities that other laboratory members

are expected to attend, such as seminars, journal clubs or team-building events.

In the first month of my six-month visiting-researcher appointment in the United Kingdom, I was invited to present my work at a monthly seminar, select an article for the journal club and join the institute's working group on research dissemination. This motivated me and made me feel like a valued member of the team.

**Assign a good workspace.** Give them a place next to or near many colleagues. This will greatly boost your new member's integration into the local team, help them to quickly grasp how things work in the new lab and spur the exchange of information and knowledge.

**Show interest.** While supervising an exchange PhD student in Belgium, I asked her how often she would like to meet with her collaborators. As a team, we then adjusted to her preference. Embrace the opportunity to get to know different styles of, for instance, project management, giving and receiving feedback and supervising PhD students.

The new colleague might have a career path that is not typical for your field in your country. Rather than wondering whether they are

## Work / Careers

equally qualified, compare what you have both learnt in the course of your careers, and consider how your experiences and skills could complement each other's.

**Talk extensively about cultural differences.** The value of international collaborations comes from the different perceptions, communication styles, work styles, customs and other forms of variety that a new colleague can bring to both the work and the social spheres. Understanding and speaking about differences will help to prevent conflicts that might arise. If they occur nonetheless, you will be able to more easily agree on how to handle similar situations in the future.

On arriving in Belgium, I noticed that I was used to a more 'direct' communication style than were my local colleagues, as a result of my previous research post in the Netherlands. This prompted me to ask the locals for advice on how to approach discussions with my PhD supervisors, interviews for grant applications or negotiations with project partners. I am convinced that such exchanges have greatly helped me to establish and uphold successful collaborations.

**Offer help with practical matters.** Navigating a new health-care system, finding a good phone plan or arranging childcare are time-consuming at best and nerve-racking at worst. I was forever grateful to a co-worker in Belgium who helped me to arrange health insurance and an annual public-transport pass and pointed out the place that sells the best bread in town.

Ideally, your university's international office should provide a welcome pack that explains which administrative and practical matters need to be arranged and how. Make sure that the new researcher gets one before their arrival.

**Pay attention to the little things.** Little things count for someone whose life – particularly their social life – has been turned upside down. Suggesting a coffee together outside work or offering to show your new colleague a nice place in town might help them to forget the stresses of relocating for a little while.

During my stay in the United Kingdom, I joined a group of fellow PhD students for weekly Friday breakfasts at a cafe close to our university. These mornings allowed us to bond, and I learnt a great deal about UK life outside the workplace. If you get along with the newcomer, you might make a new friend.

**Lara Pivodic** is a postdoctoral fellow of the Research Foundation–Flanders in Belgium and a senior researcher at Vrije Universiteit Brussel, where she conducts research on palliative and end-of-life care.

Twitter: @LaraPivodic



# DIVERSITY DEFICIT DRAGS ON

There are promising signs for gender and ethnic representation in US graduate programmes, but parity is still far off, says study. **By Virginia Gewin**

**T**he number of Indigenous and Latinx students enrolling in US graduate-level programmes for the first time rose between autumn 2017 and 2018, according to a report from the Council of Graduate Schools (CGS) in Washington DC, which represents more than 500 universities, mainly in the United States.

The report found that first-time enrolment in PhD and master's programmes grew by 8.3% and 6.8%, respectively, among American Indian/Alaska Native and Latinx students (Latinx refers to US residents with origins in

comprise 11.6%, the report found. By comparison, according to the 2010 US census, black and African American people represent 13.4% of the nation's population, and Hispanic or Latinx individuals represent 18.3%.

Indigenous students represent less than 1% of first-time enrollees, even though their total enrolment rose by 1.2% from a year ago. "More work has to be done in the graduate-education community to increase the representation of these students in the science, technology, engineering and mathematics fields," says report co-author Hironao Okahana, the CGS's associate vice-president for research and policy analysis. "All fields of study need to be a welcoming place for people from a variety of different backgrounds."

First-time international enrolment in graduate-level programmes fell for the fifth consecutive year, and is now down to 20% of all enrolment. The study found that the decline was most marked in engineering programmes, in which first-time enrolment of international students fell by 8.3% from 2017 numbers.

Female students continue to be outnumbered by their male counterparts in some graduate-level STEM programmes. They account, for example, for only 38.2% of physical and Earth-sciences graduate students and 32.1% of maths and computer-science students. "While the rate of growth for women in sciences looks good, there is still a long way to go to catch up," says Okahana.

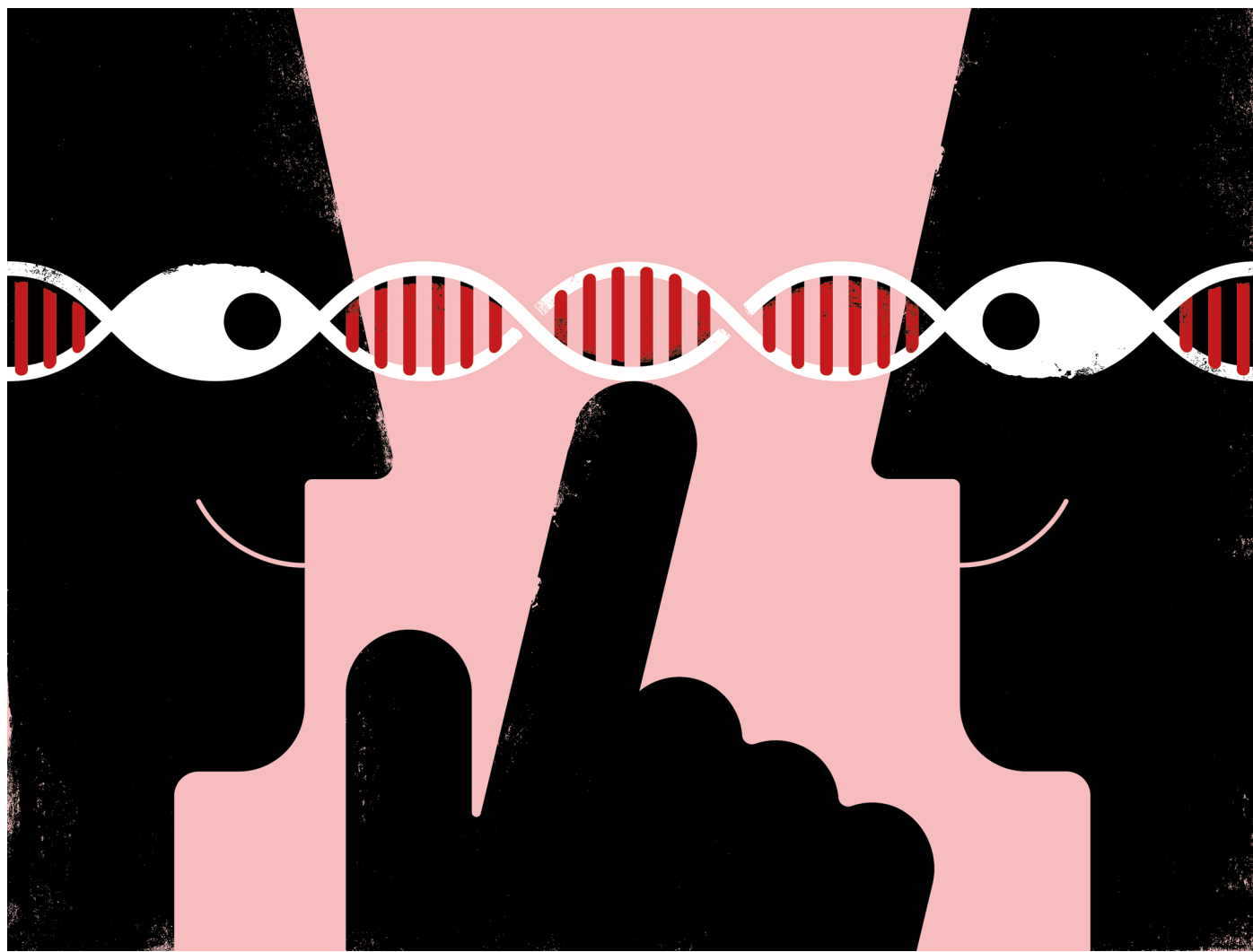
**Virginia Gewin** is a freelance writer in Portland, Oregon.

**"All fields need to be a welcoming place for people from a variety of different backgrounds."**

Latin America). Among other science, technology, engineering and mathematics (STEM) fields, maths and computer science saw a 40% and 14.2% increase in the proportions of American Indian/Alaska Native and Latinx students enrolling, respectively. The proportion of black and African American first-time enrollees in physical and Earth sciences rose by 12.5%. Results are based on responses from 589 institutions in an annual survey.

Yet, overall, US graduate-level programmes still have low proportions of students from minority ethnic groups. Black and African American students comprise 11.8% of total first-time enrollees, and Latinx students





THE PROJECT TWIN

# A PICTURE IS WORTH A THOUSAND BASE PAIRS

A small but powerful toolset makes sharing genomic data visualizations straightforward. **By Anna Nowogrodzki**

**W**hen Adam Siepel was building algorithms for evolutionary genomics as part of his PhD, he wasn't thinking about visualization. But, as a graduate student in the laboratory of computational biologist David Haussler, at the University of California, Santa Cruz (UCSC), he happened to sit next to the software engineers who were building and maintaining a tool called the UCSC Genome Browser. These engineers helped Siepel to make his algorithms publicly available as a track, or data overlay, that anyone could explore.

Genome browsers are graphical tools that display the genome sequence, usually as a horizontal line. Other sequence-associated

data are aligned and stacked above and below that line in 'tracks', for instance to illustrate the relationship between gene expression, DNA modification and protein-binding sites.

Siepel's track identifies sequences that have been retained over evolutionary time; when a user applies it while viewing the alignment of genomic data from two or more species, the track highlights regions that are evolutionarily conserved. Allowing others to use the algorithm to highlight regions of interest in their own data was "probably the single most important thing I did during my PhD", says Siepel, who is now a computational biologist at Cold Spring Harbor Laboratory in New York. Other researchers have used it, for instance, to

find mutations associated with diseases and to pinpoint functionally important regions of noncoding RNA molecules.

Today, a growing collection of free and open-source tools exists for sharing such genomic data. Which one is right for you depends on what kind of sharing you want to do: communicating with a collaborator, for instance, requires different software from what you'd use for disseminating data to the broader scientific community.

Whatever the motivation, sharing genomic data broadens its impact, says Siepel. "Almost all of our most-cited papers are supported by browser tracks," he says.

For broad dissemination of genomic data,

## Work / Technology & tools

Siepel recommends the approach that worked for him: making a track. And he suggests two genome browsers to display them: “UCSC and Ensembl are the leaders,” he says.

### Dissemination stations

The UCSC Genome Browser and its ‘track hubs’ – data tracks that are hosted remotely by external teams – can run in a web browser or on the desktop. The desktop version is called Genome Browser in a Box. Users can submit data to be included as a public track registered with UCSC. The team typically accepts only widely useful data (and not, for instance, those limited to a specific disease), and publication in a peer-reviewed journal is a plus. Alternatively, users can build personal track hubs, which involves formatting their genomic data in files of a specific format, indexing those files and making them web-accessible.

The Ensembl genome browser, hosted at the European Molecular Biology Laboratory’s European Bioinformatics Institute in Hinxton, near Cambridge, UK, allows users to import data as custom tracks, just as UCSC does; both its own format and the UCSC formats are supported. The Ensembl team has built a searchable Track Hub Registry to make it easy to find relevant track hubs for use with Ensembl or the UCSC Genome Browser.

The UCSC browser accumulates roughly two million hits a day, says Robert Kuhn, associate director of the UCSC project. And several large projects use it to disseminate data. The Genotype-tissue Expression Project, for instance, used the browser to create a track that visualizes as many as 53 tissues from 1,000 donors. The journal *Nucleic Acids Research* requires authors with whole-genome data to create a track hub for reviewers, Kuhn notes, and some authors choose to make them available to readers as well.

If you have basic Unix command-line skills, setting up a UCSC track hub takes just a few hours, says Kuhn. Instructions are available at [go.nature.com/2pqkym](http://go.nature.com/2pqkym).

### Shareable, embeddable

UCSC and Ensembl allow researchers to share data sets as tracks in a public, centralized database that is controlled by others and available to any user of that database. If your goal is to embed a visualization in a website, or to create a specialized visualization for a paper, other options are available; these include GIVE, JBrowse and IGV. (UCSC and Ensembl can also perform these tasks.)

GIVE is an open-source tool that allows researchers to build custom genome browsers for their labs with little if any programming. According to Xiaoyi Cao, a GIVE developer and a software engineer at Google, there are three ways to host data. One is for researchers to build an entire GIVE instance on their lab server using GIVE-Docker, a pre-packaged

version of GIVE that the container engine software Docker can run immediately. Because the data can remain on a private server, they do not have to be visible to the web and thus can be more secure.

Alternatively, labs can submit a list of URLs that point to the data sets they want to include, and GIVE will build the database for them, no programming required. The data can be in any of multiple formats, including those for gene-expression and protein-binding data. The resulting database will be based at a GIVE instance, or mirror, hosted by the University of California, San Diego (UCSD). And according to Sheng Zhong, the UCSD computational biologist who heads the GIVE team, it takes just two to three minutes to set up.

The third option is to include your data in the public GIVE data hub. Researchers submit their metadata to an online form, and the GIVE

**“We have all these wonderful new data types, and we have to figure out how to visualize and combine them.”**

developers will let them know if their data have been selected.

JBrowse can run either in a browser or on the desktop. Ian Holmes, JBrowse’s lead developer and a computational biologist at the University of California, Berkeley, says that he designed the tool to be responsive, intuitive and accessible for non-coders. With the desktop version, users can load data directly from their computer, Holmes says. The browser version requires an index file that tells the browser where to find the relevant data; it also requires data to be web-visible (for example, in the cloud or on a lab server). The JBrowse community has compiled a repository of about 50 plugins “that significantly enriches the visualization”, says Holmes. One example allows users to see all DNA methylation results in a single track.

Several specialized genome browsers and databases use JBrowse for data visualization; among these are the cancer genome browser COSMIC and VEuPathDB, a genomic database for pathogens and disease vectors. David Beare, a computational biologist at the Wellcome Sanger Institute in Hinxton, says that COSMIC uses JBrowse in part because “it was faster and more responsive, and certainly more intuitive” than other genome-browser options available. The VEuPathDB database developers found that JBrowse “was most amenable to our own active development” of plugins, says Omar Harb, a microbiologist at the University of Pennsylvania in Philadelphia, and director of scientific outreach and education for VEuPathDB.

JBrowse has also been used to build a

collaborative annotation tool called Web Apollo, which allows multiple researchers to simultaneously annotate the same data in real time, as in Google Docs.

### Get a Broad view

IGV is a genome browser maintained by UCSD and the Broad Institute of MIT and Harvard in Cambridge, Massachusetts. Available in desktop, browser-based and embeddable Javascript versions, IGV can generate QR codes (square barcodes) for specific data visualizations; for example, for inclusion on a poster. “IGV is always run locally on a user’s computer. There is no notion of ‘uploading’ data or saving sessions to a central IGV server hosted by us,” says Helga Thorvaldsdóttir, a software engineer at the Broad Institute. That also makes the system compatible with restricted data.

Jim Robinson, IGV’s lead developer at UCSD, says that the browser is fast and easy to use. “Most users can learn the basics in a half hour or less,” Robinson says. And the tool has racked up more than 7,000 citations, Thorvaldsdóttir says. At the Memorial Sloan Kettering Cancer Center in New York City, researchers have used IGV to visually check the genomic variants of patients whose cancer they sequence, says Robinson.

Prospective users of these tools can find plentiful educational resources online, including video tutorials. The UCSC Genome Browser has two archived and searchable listservs, or electronic mailing lists: one for website and data questions, the other for queries on setting up and maintaining Genome Browser mirrors. JBrowse users can ask questions on Github or on the software’s open instant-messaging channel, but Holmes suggests contacting the developers directly. “We have some developers who really like getting feedback from users,” he says.

And that includes suggestions for handling new challenges. Despite their utility, genome browsers are still mostly built on a fundamental assumption: that genomic data are best displayed in a linear format. But that doesn’t work so well for some kinds of information, including interactions between distant genomic regions, and evolutionary relationships, says Siepel. Some researchers, such as Maria Nattestad, a bioinformatician at Google in Palo Alto, California, have built niche tools for tackling these issues. Nattestad built a tool called Ribbon to better visualize long read alignments, for instance: these can snarl up other browsers because they often align with more than two places in genome.

“We have all these wonderful new data types, and we have to figure out how to visualize and combine them,” says Nattestad. “It keeps me up at night in the best way.”

**Anna Nowogrodzki** is a journalist in the Boston area of Massachusetts.





## Where I work Nathalie Cabrol

**I** define myself as a 'grounded mystic': that's how I feel when I am in the desert. I'm trying to understand the Universe we are living in and how to search for life in it. That's the 'grounded' aspect. The 'mystic' part is that I'm not afraid of letting my mind wander in those spaces.

I am an astrobiologist who specializes in planetary science, and my team and I study analogues of early Martian environments that might have been habitable for life as we know it. We want to understand the distribution and abundance of life in very, very harsh conditions – similar to what Mars might have been like 3.5 billion years ago – to determine how life forms survived under intense ultraviolet radiation. And we want to know how to detect and identify those life forms.

I love the barren aspect of the Altiplano in South America (this picture was taken at Salar de Pajonales in Chile, at the southwestern edge of this vast Andean plateau). In deserts, you have to be face-to-face with yourself – there is nothing else. That gives me the space I need to get thinking. There are no limitations or constraints or boundaries.

In my most recent research trip to Salar

de Pajonales this autumn, we worked more on understanding the distribution patterns of microbial life there. Life's distribution is fractal in nature and repeatable, but you have to understand the starting pattern. We are learning how to decode where to find extreme microbial life. We can then apply these codes to future missions to Mars.

In 2006, we went scuba diving in the crater lake of the Licancabur volcano, on the boundary between Chile and Bolivia. I was in completely transparent waters. The colours ranged from pale blue to dark blue, and you could see each ray of sunlight diffracted in the lake. Suddenly, it was as if the boundaries between me and that lake had completely disappeared. It was complete peace.

In science, you have to find reasons: the whys and the hows, the whats and the whens. And this moment was without time, without space. The lake was not hostile, and there was no separation between it and me.

**Nathalie Cabrol** is director of the SETI Institute's Carl Sagan Center for the Study of Life in the Universe in Mountain View, California. **Interview by Josie Glausiusz.**

Photographed by  
Andrea Frazzetta.

## Curing What Ails Us



### DOCTORS HAVE BEEN TREATING THE SYMPTOMS

of most diseases, and not the source, for centuries. They have cut out tumors, unclogged arteries, injected insulin and soothed fevers—and have been unable to touch the biological code within cells that tells them to grow malignantly, pass along abnormal nerve signals, take in too much or too little energy, and swell with inflammation. The code is the DNA molecule in each cell that tells it what to do and when, and it triggers dreaded diseases when it goes wrong.

The molecule, and its messengers, had remained tucked away, beyond the reach of almost all drugs, unfixable when broken. But as this special report explains, that is no longer the case.

Things began to change after the DNA sequence for the entire human genome was laid out early in this century, and within the past several years the ability to synthesize and custom-design shorter sequences has shown scientists that the best substance for reaching DNA is, well, DNA. Fabricating new genes to replace badly working versions, or to “silence” them, has produced 14 approved DNA-related drugs (*page S12*). And the latest research indicates that such therapies can be even more effective if scientists depart from the basic linear strands and instead make DNA spheres, which have enhanced abilities to enter cells (*page S3*). DNA analysis has also yielded new targets, showing that although newborn babies in the U.S. are typically screened for between 30 and 60 genetic conditions right now, it is possible to find nearly 1,000 genes linked to childhood diseases that could be new treatment points (*page S8*).

But that same science has also created troubling issues: some of the gene tests for infants can raise false alarms, for instance, and not every child with a disease-associated gene ends up getting that disease. Research has also revealed unfair bias in DNA targets. Most of the data about those sequences comes from studies of white people and has missed gene variants that cause disease in nonwhites—inequality in research that will produce inequality in health if it isn’t fixed (*page S14*). Geneticists are starting projects designed to improve this diversity level. DNA in medicine has great power, and that power should be used for the many, not the few.

This report on DNA drugs and related therapies, which is being published in *Scientific American* and *Nature*, is sponsored by UPMC. It was produced independently by *Scientific American* editors, who have sole responsibility for all editorial material. UPMC agreed to sponsor this topic but had no input into the content.

Josh Fischman  
Senior Editor

### S3 The Power of Spheres

DNA or RNA molecules, arranged into spherical shapes, can attack brain cancers and other illnesses that evade conventional drug design.

By Chad A. Mirkin, Christine Laramy and Kacper Skakuj

### S7 GRAPHIC: DNA TO TREAT DNA

### S8 23 and Baby

We now have the ability to screen for thousands of genetic diseases in newborns. That may not always be a healthy thing to do.

By Tanya Lewis

### S12 Gene Therapy Arrives

After false starts, drugs that manipulate the code of life are finally changing lives.

By Jim Daley

### S14 All of Us

DNA-based medicine needs more diversity to avoid harmful bias. One big research project is beginning to fix that.

By Stephanie Devaney

#### EDITORIAL

ACTING EDITOR IN CHIEF  
Curtis Brainard

CHIEF FEATURES EDITOR  
Seth Fletcher

SENIOR EDITOR  
Josh Fischman

CREATIVE DIRECTOR  
Michael Mrak

SENIOR GRAPHICS EDITOR  
Jen Christiansen

ASSOCIATE GRAPHICS EDITOR  
Amanda Montañez

COPY DIRECTOR  
Maria-Christina Keller

SENIOR COPY EDITORS  
Daniel C. Schlenoff,  
Aaron Shattuck,  
Angelique Rondeau

MANAGING PRODUCTION EDITOR  
Richard Hunt

PREPRESS AND QUALITY MANAGER  
Silvia De Santis

PUBLISHER AND VP  
Jeremy A. Abbate

DIRECTOR, INTEGRATED MEDIA  
Jay Berfas





*DRUG DEVELOPMENT*

# The Power of Spheres

DNA or RNA molecules, arranged into spherical shapes, can attack brain cancers and other illnesses that evade conventional drug design

*By Chad A. Mirkin, Christine Laramy and Kacper Skakuj*



**BRAIN CANCER IS TERRIFYING.** It attacks an organ we see as the core of our personality, our mind, our very humanity. And because the disease grows inside the brain, it is notoriously difficult to treat. The organ has evolved many defenses to keep foreign substances out as a method of self-protection, but those substances include many anticancer drugs. Using knives or radiation on this citadel of consciousness carries tremendous risks. For these reasons, the five-year relative survival rate for people aged 55 to 64 who get glioblastoma, the most common type of primary brain tumor, is a grim 5 percent. The disease killed John McCain, Edward Kennedy and Beau Biden, and it takes the lives of about 15,000 less famous Americans every year.

Now we have developed a nano-sized drug that travels through the body and into the brain, where it can kill off cancerous cells. These drug particles are composed of oligonucleotides—strands of DNA or RNA, the molecules that make up the master code that tells every cell what to do—and they stick out from a central core like the many spines of a sea urchin. The spiny round particles are called spherical nucleic acids. In an early trial with eight patients, these spheres went into glioblastoma cells and bound up other “code” molecules that are key to the cancer’s incessant growth.

Such spherical drugs appear to work against a variety of diseases. Another terrible affliction, this one affecting infants, is spinal muscular atrophy, or SMA. It robs children of muscle control until swallowing and breathing become first difficult and ultimately impossible. Most youngsters with the disorder succumb before they enter kindergarten, and until recently there was no help doctors could offer. In 2016 the U.S. Food and Drug Administration approved one remedy: a drug called Spinraza that is injected directly into the spinal cord several times every year and, at a list price of \$125,000 per shot, is one of the most expensive drugs in the world. We recently compared our spheres, studded with nucleic acids that get inside cells and interfere with messenger molecules that lead to SMA’s symptoms, with the Spinraza approach in studies of rodents. The spheres improved survival by four times—115 days versus 28 days—and the rate of toxic side effects was much lower.

Spherical nucleic acids, or SNAs, avoid problems that have plagued the pharmaceutical industry’s attempts to develop new drugs. Conventional drugs are nonspecific: they can affect many cells and organs, not just diseased ones; hence, they have numerous side effects. Nucleic acids, however, can be designed to interfere with only disease-causing genes or their related instruction molecules sent to control a cell’s behavior. Biologists have tried to use nucleic acids in the past but primarily as linear molecules and with little ability to direct where they go. And because the body has robust defenses against foreign genetic material—the immune system, for one—in most cases, these defenses damaged the drugs immediately or sent them to organs such as the liver and kidneys for waste removal.

But SNAs, at only billionths of a meter across, seem able to travel anywhere in the body and get inside cells before immune defenses can waylay them. The spherical shape lets us pack a high density of nucleic acid “spines” into a small space, and that density creates a strong interaction with receptors on cell surfaces that admit the particles inside. There the sequence of the components—the same nucleotides, abbreviated as A, T, C and G, that constitute the DNA

code of life—ensures that they affect only complementary sequences of DNA or RNA. (The latter molecule uses U—uracil—instead of T, and we design for that.) We construct our strands to match only sequences in the cells that are crucial to the disease. SNAs are not magic bullets and will have to pass many more tests before they can be used on lots of patients. But the potential is there: because the nucleic components can be reordered to interfere with many different disease-causing molecules within cells, the spheres have the ability to tackle some of the world’s most debilitating conditions.

#### PROGRAMMABLE DRUGS

**TRADITIONALLY, SCIENTISTS** have found disease treatments by screening hundreds of thousands of small synthetic or natural molecules, going through a long trial-and-error process to see if any of them have therapeutic benefits. Although this pipeline has led to a number of amazing medicines, such as antibiotics, even the most promising ones can cause unwanted side effects. Many other diseases are unaffected by these molecules and therefore still lack a cure or treatment. Even biologics, a newer class of drugs that are often based on proteins made by immune cells of mice, rabbits and other animals, typically rely on an abbreviated trial-and-error discovery process.

An ideal drug-design process would allow scientists to rapidly and rationally design specific drugs that use the same language as our cells, instead of looking for a needle-in-a-haystack molecule. Cells communicate many complex messages through DNA and RNA to make millions of proteins. The number of steps that cells must execute correctly to make these proteins is staggering: they must select a specific sequence of DNA made of A, T, C and G nucleotides, transcribe that sequence into a form called messenger RNA (mRNA), and then accurately read that mRNA to arrange molecules called amino acids into a chain—as long as 35,000 units—that forms a single protein.

Errors where one nucleotide such as a T or a G is added, deleted or placed in an incorrect order can halt protein production or generate an irregular protein that causes disease. Too many copies of an mRNA, and therefore of its related protein, can also lead to disease. (So can the introduction of foreign nucleic acids from a virus, which leads the infected cell to make harmful viral protein.)

But we can synthesize our own stretches of DNA or RNA components, called oligonucleotides. Because the genetic alphabet has very specific rules—A can bind only to T, and C binds only to G—we can make our oligonucleotides with sequences that selectively bind to and inactivate one disease-driving sequence. When they do so, the synthetic oligonucleotides gum up the cellular works, pre-

venting the affected cells from producing a disease-causing protein.

Yet despite automated equipment that can rapidly make synthetic oligonucleotides with any desired sequence one could imagine, fewer than a dozen oligonucleotide-based drugs have been approved for patients. This is because these strands of oligonucleotides face a significant hurdle once they are injected into the bloodstream: because they are foreign—that is, not native to the patient—they get treated as hazardous material or waste. The body's immune system either destroys these oligonucleotides, or the body's waste-filtration stations, the liver and kidneys, remove them. They do not reach their intended target. Even if oligonucleotide strands could make it to a cell that contained the target mRNA, that cell has an outer membrane that acts as a barrier to prevent the oligonucleotides from getting inside. As a result, drug companies working with oligonucleotides have often settled for treating diseases that can be targeted in the liver. The liver is an important organ. But sequestering these drugs in this one place really limits their use. (An alternative approach—injecting oligonucleotides directly into the disease site, such as into the spinal column with Spinraza—is technically difficult and still does not ensure entry of the medicine into all the appropriate cells.)

#### A SURPRISING RESULT

**ADVANCES IN NANOTECHNOLOGY** made by our group at Northwestern University, along with several other researchers, have led us to the SNAs, which may be a way around this problem. Prior to 2006, our group had been interested in using the highly specific binding ability of SNAs in probes for ultra-sensitive diagnostics—to fish out stretches of cancer DNA from blood samples, for instance. We could do this by chemically decorating a gold nanoparticle with many strands of DNA designed to anchor one end to the particle, producing the sea urchin spine pattern. The outer end of the DNA was designed to be a complementary sequence to the cancer DNA sequence, so it worked nicely as a probe. We also used the spheres as artificial atoms with programmable bonds to fashion new types of materials. Drug design, however, was not really on our radar. After all, according to the dominant paradigm of drug biology and chemistry, RNA and DNA would not naturally cross cell membranes.

We were curious, though, about how nucleic acids in this new geometry would interact with living systems. Drug developers had already been experimenting with single strands of oligonucleotides, with, as we noted, limited success. From our research with SNAs as a diagnostic platform, we knew that target DNA and RNA would bind to our clusters of spines much more strongly than they would attach to free oligonucleotide strands. The reason is that our spines are packed densely on the nanoparticle's surface. That makes them more rigid, which helps the As, Ts, Gs and Cs on each strand align and bind when they encounter a target strand. This characteristic made us suspect that with the right nucleic acid sequences, SNAs could be a very potent oligonucleotide drug.

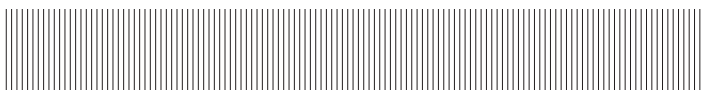
To test this idea, we carried out an experiment that, at the time,

we thought had only a slim chance of working. We took strands of free oligonucleotides and put them into a test tube with mouse cells. In a different tube we added a bunch of SNAs to the same type of mouse cells. We attached red fluorescent molecules to both the spheres and the strands to help us track them. When we looked at the cells under a microscope, the ones mixed with free strands appeared transparent, as expected. Free oligonucleotides did not cross the cell membrane. But the cells mixed with SNAs lit up the screen with bright red fluorescence. The spheres had made it inside!

How could this happen? In general, cell membranes closely regulate which molecules may enter, and oligonucleotides are not typically among the approved guests. Furthermore, oligonucleotides carry a negative electrical charge, as do cell surfaces. Like two magnets, the two biological objects should repel each other. Yet when we repeated this experiment over and over again using more than 50 other human and animal cell types, all but one glowed red, a signal of success.

Today we think we know what the gateway is: a type of doorway

### The ability of SNAs to reach the brain and their lack of toxicity generate hope for treating a dangerous cancer, as well as other neurological disorders, and set the stage for the next set of clinical trials.



molecule called a scavenger receptor that dots the cell surface. These receptors play a major role when a cell engages with its environment; for example, they admit nano-sized biomolecules the cell needs. Some of the structural features at the ends of SNA spines happen to mimic the natural substrates of these scavenger receptors. As noted earlier, the strands on the spheres are densely packed, and like with Velcro, the more hooks, the stronger the bond. With free strands, even if scavenger receptors recognized them as molecules to take in, they have only one hook and float away.

With the aid of an electron microscope, we could see that once an SNA binds to these receptors, the surrounding cell membrane folds inward to create a pocket, ushering the SNA into the cell.

#### SPHERES AS MEDICINE

**BUT GETTING IN** was only half the battle. To work as a drug, the SNA needed to find, bind to and inactivate a particular stretch of mRNA that instructed the cell to make a disease-associated protein.

The first stretch of mRNA in a cell that we targeted did not cause disease but did instruct the cell to make a protein that glowed bright green under a microscope. Our goal was to stop this mRNA. When

we exposed mouse cells to an SNA designed to match that green-causing mRNA and compared them with similar cells that did not get the spheres, the color difference was clear. Sphere-free cells were bright green, showing the mRNA had encoded proteins. But cells exposed to our SNAs were transparent, meaning we had blocked the mRNA before it could pass along instructions to make anything green, as we reported in *Science* in 2006.

Next we pitted SNAs against the major challenge plaguing linear oligonucleotide drugs: destruction by the body's natural defense system. We found that our spheres have a strong electrical charge—again because of the dense packing—that helped them evade immune interference. This high charge inhibits defense molecules called nucleases, proteins that degrade foreign DNA and RNA, from getting close.

### REALITY TEST

**WE WERE ON TO SOMETHING**, at least in the laboratory. Other scientists replicated and independently advanced some of our work, including dermatologist Amy Paller, Arthur Burghes, an expert on SMA, immunotherapy specialist Bin Zhang, cancer biologist Alex Stegh, transplant surgeon Jason Wertheim, and oncologist Priya Kumthekar. But the path from benchtop breakthroughs to healthier patients is long and hard, so nearly 10 years ago researchers from our group founded a company called Exicure to advance SNA-based drugs to the clinic.

We initially explored whether these potent drugs could be delivered to diseased tissues in skin creams and eye drops, which is feasible because SNAs are easily taken up by cells and a big improvement over invasive strategies such as direct injections. Two of our first targets were psoriasis and poorly healing wounds, and there are several promising SNA candidates already in early-stage clinical trials for some of these ailments.

Skin, of course, is relatively easy to get to. The brain is not. Defended by a vigilant immune system and a web of blood vessels—the blood-brain barrier—designed to keep foreign molecules out, the brain makes cancers such as glioblastoma particularly difficult to treat. We thought, however, that SNAs might move across these defenses via the same doorway molecules that ease their path through cell membranes. Once in the brain the spheres could home in on cancer cells by targeting genes and proteins responsible for keeping the cells alive, which malignancies produce in excessive amounts.

To start this project, we created an SNA drug with many short pieces of RNA specifically designed to knock down the production of a protein in glioblastoma cells called Bcl2L12. That protein acts as a biochemical defender that helps to keep the cancer cells functioning. We thought that by intercepting the mRNA that tells the cells to make this protein, the SNAs could make the cancer vulnerable to conventional medicines. Indeed, in our animal studies, reported in 2013 in *Science Translational Medicine*, that is what happened: SNAs injected into the bloodstream of mice reached the brain, crossed the blood-brain barrier and prevented the production of Bcl2L12 protein inside of glioblastoma cells. Last year early clinical results showed that these SNAs also reach glioblastoma cells in human patients. We did not cure people, and we have yet to test whether the SNAs make the cancer cells more vulnerable. Still, the ability of SNAs to reach the brain and their lack of toxicity generate hope for

treating this cancer, as well as other neurological disorders, and set the stage for the next set of clinical trials. And tests in other diseases, such as spinal muscular atrophy, show promise in animals.

Another exciting direction for SNAs is their use as immunotherapies against cancer. Cancer cells often have proteins in their membrane that are different from the proteins found in healthy cells. Therefore, a cancer cell protein can act as a red flag, and if our immune system can be trained to go after it the way it goes after a flu virus, our own bodies can do a better job of protecting us from the disease.

To make an SNA cancer vaccine, we exchanged the gold-nanoparticle core for a hollow nanoparticle called a liposome, filled it with one of these red-flag proteins and injected it into animals with the corresponding cancer. Some of our most recent experiments, published in 2019 in the *Proceedings of the National Academy of Sciences USA*, showed that such SNAs elicit an immediate immune response to the tumor, apparently teaching the immune system to go after cells showing that red flag. The effects appear long-lasting, too: the immune system keeps going after cells with that protein after the SNAs have vanished. SNAs are already showing potency and safety in phase I clinical trials in humans, and other spheres targeting a deadly skin cancer are being tested in a separate set of safety trials.

SNAs are, however, not yet approved drugs. There are a number of challenges that they have to overcome first. Because the spheres do get to a wide set of cells, we need to carefully study whether or not they produce any negative “off target” effects even though their design should limit them to only problem DNA and RNA. Larger patient populations must be explored, and we need to improve targeting to increase the amount of drug that gets to the affected organ and cells.

We think the ability of SNAs to access so many different tissues is game-changing and will be central to the emergence and ultimate widespread use of such medicines. SNAs are the product of three core capabilities: the ability to make large quantities of oligonucleotides, an understanding of genetic disease pathways, and the ability to get such oligonucleotides into tissues and cells that matter. The first two are important, but without the third the process is like making software without hardware it needs to run on. SNAs may be that crucial and versatile hardware—a platform able to be reused for many different types of illness, one that begins to move the pharma industry away from the difficult search for entirely new molecules for every new treatment. An SNA simply needs a different set of oligonucleotides to be sent after a new disease. And we are just getting started.

---

**Chad A. Mirkin** is director of the International Institute for Nanotechnology and holds professorships in chemistry, chemical and biological engineering, biomedical engineering, materials science and medicine at Northwestern University. He is a founder of Exicure, a company developing spherical nucleic acids for use as drugs.

---

**Christine Laramy** received her Ph.D. in chemical and biological engineering from Northwestern and is now an analyst at the law firm Latham and Watkins.

---

**Kacper Skakuj** is a graduate student in the chemistry department at Northwestern.

---



# DNA to Treat DNA

Within a cell, aberrant DNA—and the messenger RNA (mRNA) it uses to tell the cell what to do—can cause disease. Scientists can synthesize DNA that specifically binds to such problem molecules. When formed into spherical nucleic acids (SNAs), it penetrates cells and interferes with the trouble-causing molecules.

## LINEAR LIMITS

DNA or RNA drugs have been tried with the more typical linear strands of the molecules. These can work but often have difficulty entering a cell or are destroyed by immune defenses. They usually need to be injected directly into a disease site, which limits use.

Linear form (oligonucleotide)

Target cell

## DISEASE CONTINUES

Molecules of mRNA associated with disease, which instruct the cell to make proteins, are unimpeded.

Unwanted proteins

Traditional small molecule drugs target the resulting proteins

SNAs start with a core, often made from a nanoparticle called a liposome. Custom-made single-stranded DNA is packed densely around that core.

Nanoparticle core

Anchor

Single-stranded DNA

## SPHERICAL SUCCESS

On the surface of the SNAs, the many strands of DNA show abundant attraction points to cell doorways called scavenger receptors, in contrast to the single “hook” of a free strand. Thus, the spheres are more easily taken inside the cell.

Objects not drawn to scale

Scavenger receptors

Captured mRNA

## ILLNESS INTERRUPTED

The custom sequences on SNAs bind only to the mRNAs involved in disease, ignoring other molecules. Once captured in this way, the mRNAs can no longer tell the cell to make proteins that harm the body.

Fewer unwanted proteins







## MEDICAL TESTS

# 23 and Baby

We now have the ability to screen for thousands of genetic diseases in newborns. That may not always be the healthy thing to do

By Tanya Lewis

**MITCHELL GORBY CAME INTO** this world around 3 P.M. on August 9, 2019, at Balboa Naval Hospital in San Diego. The baby seemed healthy, and his parents, Tiffany and Rylan, were thrilled. But a few hours later a nurse noticed that Mitchell seemed lethargic and never cried, and monitors indicated that his body was not getting enough oxygen. Mitchell was rushed to the neonatal intensive care unit at nearby Rady Children's Hospital, where tests revealed that oxygen wasn't bonding to the molecule that carries it through the blood, hemoglobin, and his red blood cells were dying off. He wasn't nursing, so the hospital put in a feeding tube. Mitchell's doctor ordered CT and brain scans and tested for infectious diseases—but she could not figure out what was wrong with him. As a last resort, she suggested sequencing Mitchell's genome.

The results from Stephen Kingsmore's laboratory at the Rady Children's Institute for Genomic Medicine came back within about 48 hours. Mitchell had a rare genetic mutation known as hemoglobin Toms River, which prevents oxygen from bonding to the proteins in fetal red blood cells. The mutation—named after the New Jersey hometown of the first patient identified with the problem in 2011—affects only fetal hemoglobin; babies start making healthy adult hemoglobin within a few months. Doctors just had to keep Mitchell alive until that happened. Rady neonatologist Jeanne Carroll says that “having his whole genome allowed us to know the starting point” for treatment. She and Mitchell's team of physicians prescribed a series of blood transfusions, and the baby improved rapidly. In just under a month he was strong enough to go home.

For children like Mitchell who are born with a genetic disease, it used to take years to get a diagnosis, and by then it often was too late. Now, however, advances in the speed of genetic sequencing and steeply falling costs have made it possible to screen for hundreds or even thousands of childhood-onset genetic diseases. Within the past year or so a few dozen hospitals have started offering the ability to rapidly sequence a newborn's genome to help diagnose a life-threatening condition soon after birth. Researchers are studying whether such sequencing should be offered to all newborns as part of standard health screening. And companies such as Sema4 and BabyGenes are now marketing 23andMe-style direct-to-consumer tests to parents simply seeking to know more about the health of their baby. Prenatal and newborn genetic sequencing is expected to grow to an \$11.2-billion industry by 2027, up from a \$4-billion market in 2018.

Proponents say that genetic testing of newborns can help diagnose a life-threatening childhood-onset disease in urgent cases and could dramatically increase the number of genetic conditions all babies are screened for at birth, enabling earlier diagnosis and treatment. It could also inform parents of conditions they could pass on to future children or of their own risk of adult-onset diseases. Genetic testing could detect hundreds or even thousands of diseases, an order of magnitude more than current heel-stick blood tests—which all babies born in the U.S. undergo at birth—or confirm results from such a test.

But others caution that genetic tests may do more harm than good. They could miss some diseases that heel-stick testing can detect and produce false positives for others, causing anxiety and leading to unnecessary follow-up testing. Sequencing children's DNA also raises issues of consent and the prospect of genetic discrimination.

Regardless of these concerns, newborn genetic testing is already here, and it is likely to become only more common. But is the technology sophisticated enough to be truly useful for most babies? And are families—and society—ready for that information?

**IN THE 1960S MICROBIOLOGIST** Robert Guthrie developed a test for phenylketonuria (PKU), a genetic disorder that causes the amino acid phenylalanine to build up in the body. PKU is easily treated with a phenylalanine-restricted diet, but without intervention it can cause brain damage and mental disabilities. Within a few years other U.S. states required that Guthrie's test be administered to newborns, and tests for other conditions were soon to follow. By the mid-1980s most states had mandatory screening programs. In 2002 the federal government asked

the American College of Medical Genetics to develop guidelines for newborn screening, which culminated in the Recommended Universal Screening Panel, a set of 35 core conditions and 25 secondary ones that are treatable. Most states now test for a subset of these conditions.

There are roughly 14,000 known genetic diseases in humans, ranging from childhood-onset diseases such as PKU and congenital heart disease to adult-onset conditions such as Huntington's disease and heritable forms of cancer. Some childhood diseases, such as PKU, are treatable if caught early. Heel-stick tests look for only a tiny fraction of these diseases, hence the appeal of genetic testing.

In the early 2010s researchers at the National Institute of Child Health and Human Development and the National Human Genome Research Institute launched a program, called NSIGHT (short for Newborn Sequencing in Genomic Medicine and Public Health), to explore the risks and benefits of DNA screening of newborns. Rady's Kingsmore led one of four projects funded by NSIGHT, which explored the use of rapid, whole-genome sequencing in extremely sick newborns suspected of having a genetic disease.

Standard sequencing can take weeks, but using a rapid sequencing method and software that compared the genome with the patient's disease characteristics, Kingsmore's team could get a genetic diagnosis back in as little as a day or two. For these babies, hours or days can be the difference between life and death or severe disability. The first of two trials led by Kingsmore took place from 2014 to 2016 at Children's Mercy Hospital in Kansas City. The second ran from 2017 to 2019 at Rady Children's. Within the past year the group has started offering newborn sequencing at 23 hospitals around the country, and lawmakers from California have introduced federal legislation to cover the cost of sequencing critically ill babies through Medicaid. As of last November, Kingsmore and his colleagues had sequenced more than 1,100 babies with suspected genetic diseases. About one in three of them received a diagnosis that identified an illness, and one in four had their existing treatment changed as a result.

Mitchell Gorby was one of those sequenced at Rady (but not as part of NSIGHT). Carroll, the Rady neonatologist, says the information "helped us more confidently give him more transfusions and hold off on other testing." It is possible Mitchell may have survived and outgrown his disorder without the test and diagnosis. But in other cases, sequencing has very likely saved lives. Moreover, sequencing probably significantly reduced the diagnostic odyssey such children have to take, Kingsmore says.

**EXTREMELY SICK BABIES** are not the only ones who could benefit from genetic testing. Another NSIGHT project investigated whether sequencing could also be used in clinical settings to screen newborns with no obvious signs of disease.

For this study, called the BabySeq Project, Robert Green of Brigham and Women's Hospital, Alan Beggs of Harvard Medical School and their colleagues recruited families and randomly assigned half of them to have their babies' genomes sequenced. They developed a list of about 1,500 genes that were highly associated with diseases that begin in childhood or adolescence, then returned information about a subset of those genes to the families. The goal was to do the most

# 35

**genetic diseases can be spotted by blood tests for newborns used in many states. The tests look for parts of proteins or other molecules linked to treatable gene-associated ailments.**

# 193

**illnesses can now be identified through DNA itself, using one of the more popular commercial genetic test panels for newborns, Sema4's Natalis. Like state blood tests, Natalis screens for diseases that are treatable.**

# 1,514

**genes, each responsible for a different childhood disease, were identified in a research study on newborns called BabySeq. It looked for DNA tied to treatable illnesses, for genes that can affect responses to drugs, and for genes that would not affect the particular baby but could be passed on and cause disease in future generations.**

comprehensive testing possible—to see anything and everything that could be discovered about gene-based risks. Last January the group reported sequencing results from 159 newborns—mostly healthy babies but also some ill ones in the neonatal ICU. The scientists found that 9.4 percent of the healthy group were at risk of developing a childhood-onset disease that was not known from their medical or family history, and 88 percent were carriers for recessive diseases.

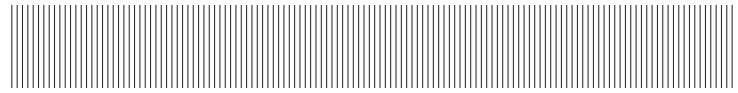
So was the testing worth it for parents? A mother named Natalie, who requested we use only her first name out of concern for her family's privacy, has a son who was enrolled in BabySeq. Natalie, who is a physician and lives in Washington, D.C., admits she felt some nervousness about the testing. "Whenever you have the chance to learn about the health of your child, there's an opportunity for anxiety," she says. But overall, she and her husband were comfortable with the project. "Because they were looking at only genetic defects that affect childhood and only illnesses that had some preventive measures, we felt it could potentially be useful," she says.

Fortunately, the results of tests on her son, Russell, did not turn up any childhood-onset genetic disorders. The exams did indicate that he may be a carrier for a recessive metabolic disorder called Gaucher disease, but the sequencing of this gene is particularly prone to error, so he will need follow-up testing to confirm. For other families, the benefits of sequencing were more clear-cut: one child had a disorder—missed by standard screening—that makes the body unable to recycle a vitamin called biotin; the condition can cause coma and death if left untreated, but it can easily be treated by supplementation.

Although BabySeq was initially focused only on childhood-onset disorders, one baby in the study was found to carry a variant of the *BRCA2* gene, which is associated with a high risk of breast and other cancers, so the researchers asked parents for permission to inform them of the risk of adult-onset disorders if they chose. Natalie and her husband opted not to receive this information but said they would leave it up to Russell if he wanted to be tested when he was older. "We felt it should be our son's decision," Natalie says.

**BECAUSE OF ITS COMPLEXITY** and cost, BabySeq was never intended to be a feasible addition to standard newborn screening. "We have not tried to advocate for this in clinical practice," Green of Brigham and Women's says. But sequencing tests are no longer confined to clinical practice. Several companies now offer direct-to-consumer DNA tests for newborns. The firm Sema4 sells a test for \$379 that it says screens for more than 190 genetic conditions that can occur before the age of 10 and that can be treated with medication, diet or other interventions. The company gives results to parents in a genetic-counseling session about four to six weeks after the test. Sema4's CEO, Eric Schadt, says the test can detect disease-related genetic variants with 99 percent accuracy. Sema4 only reports results for diseases that have a greater than 80 percent penetrance—the proportion of people with a genetic variant who end up developing the disease. It also discloses information about the child's sensitivity to certain drugs, although the U.S. Food and Drug Administration has recently been pressuring companies not to

**"Whenever you have the chance to learn about the health of your child, there's an opportunity for anxiety."  
—Natalie, BabySeq parent**



make such information available, because it says that it has not reviewed the tests and that they may not be backed up by clinical evidence.

Another company, BabyGenes, offers a test that scours 100 genes for more than 72 conditions. It is offered in the form of either a cheek swab or dried-blood spot test and retails for \$349.

Schadt admits Sema4 doesn't know whether the kind of testing it offers leads to an overall benefit for patients, although he says the company is doing studies to find out. There are reasons to wonder. The accuracy of these tests in detecting disease is still uncertain. In a third NSIGHT project, led by Jennifer Puck, Barbara Koenig and Pui-Yan Kwok of the University of California, San Francisco, researchers sequenced the DNA of dried spots of blood left over from newborn heel-stick tests (California has kept all its blood spots since the early 1980s). Although the sequencing did detect some genetic conditions that the standard newborn screening panel does not test for, it missed some of those that standard screening caught. And it flagged a lot of genetic variants of unknown significance, Puck says: "Newborn screening is very different from having a sick individual in front of you for whom you're trying to arrive at a diagnosis."

When combined with the standard screening, DNA testing did reduce the number of false positives, however. Puck thinks sequencing could be an add-on to standard screening when there's an abnormal result, but she doesn't think it should be used to screen all healthy babies. "We're just not at the point where we can interpret the sequence with sufficient predictive value to say 'yes' or 'no,' this is a disease or not," she says.

Another issue that concerns physicians and medical ethicists is the possibility that genetic testing will cause unnecessary anxiety for parents about diseases that may appear later in life or never show up at all. "When it comes to genetic information about your child, a lot of people aren't in a position to well interpret what the results mean," says Nita Farahany, a professor of law and philosophy at Duke University School of Law, who is an expert in genetics and bioethics. "If they're told their child has a four times greater risk [of some condition], but the population risk is 1 percent, how do they treat their children?" There is already a shortage of genetic counselors in the U.S., so there would not be enough people to help parents understand their child's genetic results.

Then there's the issue of privacy. If the child's genetic information is stored on file, who has access to it? If the information becomes public, it could lead to discrimination by employers or insurance companies. The Genetic Information Nondiscrimination Act (GINA), passed in 2008, prohibits such discrimination. But GINA does not

apply to employers with fewer than 15 employees and does not cover insurance for long-term care, life or disability. It also does not apply to people employed and insured by the military's Tricare system, such as Rylan Gorby. When his son's genome was sequenced, researchers also obtained permission to sequence Rylan's genome, to determine if he was a carrier for the rare hemoglobin condition. Because it manifests itself only in childhood, Gorby decided taking the test was worth the risk of possible discrimination.

Cost is another consideration. Clinical sequencing is still about \$500 to \$800, and interpretation can be upward of \$1,000, according to Brigham and Women's Green. For families who can't afford health insurance, this is out of reach. Some experts have also raised concerns that genetic testing could lead to a lot of follow-up testing with specialists, which could overburden an already resource-strapped health care system. If sequencing turns out to save money in the long run, insurance companies may cover it, but there's no guarantee.

Yet another problem is that the majority of the sequencing to date has been done in babies whose families are well-off and white, raising concerns that this could become the province of only the privileged. And the racial homogeneity could skew the results: diseases more prevalent in Caucasian individuals could be overrepresented in test panels, whereas illnesses more common in racial minorities may be underrepresented. (New medical data projects intend to address this disparity [see "All of Us," on page S14].)

**THE U.C.S.F. NSIGHT PROJECT** included a working group that investigated some of these ethical and policy issues, which culminated in a 2018 report by the Hastings Center, a bioethics nonprofit in Garrison, N.Y. The report concluded that newborn sequencing has many benefits in helping diagnose sick babies and could expand the number of conditions that meet the stringent newborn screening criteria. But using genome sequencing as a replacement for newborn screening is "at best premature," the authors say, and direct-to-consumer sequencing should not be used for diagnosis or screening purposes.

Barbara Koenig, a professor of medical anthropology and bioethics at U.C.S.F. and one of the report's co-authors, underscores the fact that sequencing, while promising, is not yet mature enough to be routinely used to screen healthy children. "This is not a technology that's ready for prime time for use in healthy infants," Koenig says.

Despite these concerns, the era of newborn sequencing is now upon us, and the practice will likely become more widespread as costs come down and the results become more accurate and useful. In the meantime, the risks and benefits of sequencing must be weighed on an individual basis. Extremely sick newborns are a completely different case from apparently healthy children of worried parents susceptible to marketing from genetic-testing firms.

For Mitchell Gorby, sequencing was certainly worth it. Two months after leaving the hospital, he is doing fine and has doubled his weight. His parents are settling into their new routine, somewhat sleep-deprived, but happy to be home with their healthy baby boy.

**Tanya Lewis** is an associate editor who covers health and medicine at *Scientific American*.

## Gene Therapy Arrives

**After false starts, drugs that manipulate the code of life are finally changing lives**

By Jim Daley

The idea for gene therapy—a type of DNA-based medicine that inserts a healthy gene into cells to replace a mutated, disease-causing variant—was first published in 1972. After decades of disputed results, treatment failures and some deaths in experimental trials, the first gene therapy drug, for a type of skin cancer, was approved in China in 2003. The rest of the world was not easily convinced of the benefits, however, and it was not until 2017 that the U.S. approved one of these medicines. Since then, the pace of approvals has accelerated quickly. At least nine gene therapies have been approved for certain kinds of cancer, some viral infections and a few inherited disorders. A related drug type interferes with faulty genes by using stretches of DNA or RNA to hinder their workings. After nearly half a century, the concept of genetic medicine has become a reality.

### GENE INSERTION

These treatments use a harmless virus to carry a good gene into cells, where the virus inserts it into the existing genome, canceling the effects of harmful mutations in another gene.

**GENCICINE:** China's regulatory agency approved the world's first commercially available gene therapy in 2003 to treat head and neck squamous cell carcinoma, a form of skin cancer. Gencicine is a virus engineered to carry a gene that has instructions for making a tumor-fighting protein. The virus introduces the gene into tumor cells, causing them to increase the expression of tumor-suppressing genes and immune response factors. The drug is still awaiting FDA approval.

**GLYBERA:** The first gene therapy to be approved in the European Union treated lipoprotein lipase deficiency (LPLD), a rare inherited disorder that can cause severe pancreatitis. The drug inserted the gene for lipoprotein lipase into muscle cells. But because LPLD occurs in so few patients, the drug was unprofitable. By 2017

its manufacturer declined to renew its marketing authorization; Glybera is no longer on the market.

**IMLYGIC:** The drug was approved in China, the U.S. and the E.U. to treat melanoma in patients who have recurring skin lesions following initial surgery. Imlygic is a modified genetic therapy inserted directly into tumors with a viral vector, where the gene replicates and produces a protein that stimulates an immune response to kill cancer cells.

**KYMRIAH:** Developed for patients with B cell lymphoblastic leukemia, a type of cancer that affects white blood cells in children and young adults, Kymriah was approved by the FDA in 2017 and the E.U. in 2018. It works by introducing a new gene into a patient's own T cells that enables them to find and kill cancer cells.



apply to employers with fewer than 15 employees and does not cover insurance for long-term care, life or disability. It also does not apply to people employed and insured by the military's Tricare system, such as Rylan Gorby. When his son's genome was sequenced, researchers also obtained permission to sequence Rylan's genome, to determine if he was a carrier for the rare hemoglobin condition. Because it manifests itself only in childhood, Gorby decided taking the test was worth the risk of possible discrimination.

Cost is another consideration. Clinical sequencing is still about \$500 to \$800, and interpretation can be upward of \$1,000, according to Brigham and Women's Green. For families who can't afford health insurance, this is out of reach. Some experts have also raised concerns that genetic testing could lead to a lot of follow-up testing with specialists, which could overburden an already resource-strapped health care system. If sequencing turns out to save money in the long run, insurance companies may cover it, but there's no guarantee.

Yet another problem is that the majority of the sequencing to date has been done in babies whose families are well-off and white, raising concerns that this could become the province of only the privileged. And the racial homogeneity could skew the results: diseases more prevalent in Caucasian individuals could be overrepresented in test panels, whereas illnesses more common in racial minorities may be underrepresented. (New medical data projects intend to address this disparity [see "All of Us," on page S14].)

**THE U.C.S.F. NSIGHT PROJECT** included a working group that investigated some of these ethical and policy issues, which culminated in a 2018 report by the Hastings Center, a bioethics nonprofit in Garrison, N.Y. The report concluded that newborn sequencing has many benefits in helping diagnose sick babies and could expand the number of conditions that meet the stringent newborn screening criteria. But using genome sequencing as a replacement for newborn screening is "at best premature," the authors say, and direct-to-consumer sequencing should not be used for diagnosis or screening purposes.

Barbara Koenig, a professor of medical anthropology and bioethics at U.C.S.F. and one of the report's co-authors, underscores the fact that sequencing, while promising, is not yet mature enough to be routinely used to screen healthy children. "This is not a technology that's ready for prime time for use in healthy infants," Koenig says.

Despite these concerns, the era of newborn sequencing is now upon us, and the practice will likely become more widespread as costs come down and the results become more accurate and useful. In the meantime, the risks and benefits of sequencing must be weighed on an individual basis. Extremely sick newborns are a completely different case from apparently healthy children of worried parents susceptible to marketing from genetic-testing firms.

For Mitchell Gorby, sequencing was certainly worth it. Two months after leaving the hospital, he is doing fine and has doubled his weight. His parents are settling into their new routine, somewhat sleep-deprived, but happy to be home with their healthy baby boy.

**Tanya Lewis** is an associate editor who covers health and medicine at *Scientific American*.

## Gene Therapy Arrives

**After false starts, drugs that manipulate the code of life are finally changing lives**

By Jim Daley

The idea for gene therapy—a type of DNA-based medicine that inserts a healthy gene into cells to replace a mutated, disease-causing variant—was first published in 1972. After decades of disputed results, treatment failures and some deaths in experimental trials, the first gene therapy drug, for a type of skin cancer, was approved in China in 2003. The rest of the world was not easily convinced of the benefits, however, and it was not until 2017 that the U.S. approved one of these medicines. Since then, the pace of approvals has accelerated quickly. At least nine gene therapies have been approved for certain kinds of cancer, some viral infections and a few inherited disorders. A related drug type interferes with faulty genes by using stretches of DNA or RNA to hinder their workings. After nearly half a century, the concept of genetic medicine has become a reality.

### GENE INSERTION

These treatments use a harmless virus to carry a good gene into cells, where the virus inserts it into the existing genome, canceling the effects of harmful mutations in another gene.

**GENDICINE:** China's regulatory agency approved the world's first commercially available gene therapy in 2003 to treat head and neck squamous cell carcinoma, a form of skin cancer. Gendicine is a virus engineered to carry a gene that has instructions for making a tumor-fighting protein. The virus introduces the gene into tumor cells, causing them to increase the expression of tumor-suppressing genes and immune response factors. The drug is still awaiting FDA approval.

**GLYBERA:** The first gene therapy to be approved in the European Union treated lipoprotein lipase deficiency (LPLD), a rare inherited disorder that can cause severe pancreatitis. The drug inserted the gene for lipoprotein lipase into muscle cells. But because LPLD occurs in so few patients, the drug was unprofitable. By 2017

its manufacturer declined to renew its marketing authorization; Glybera is no longer on the market.

**IMLYGIC:** The drug was approved in China, the U.S. and the E.U. to treat melanoma in patients who have recurring skin lesions following initial surgery. Imlygic is a modified genetic therapy inserted directly into tumors with a viral vector, where the gene replicates and produces a protein that stimulates an immune response to kill cancer cells.

**KYMRIAH:** Developed for patients with B cell lymphoblastic leukemia, a type of cancer that affects white blood cells in children and young adults, Kymriah was approved by the FDA in 2017 and the E.U. in 2018. It works by introducing a new gene into a patient's own T cells that enables them to find and kill cancer cells.



**LUXTURNA:** The drug was approved by the FDA in 2017 and in the E.U. in 2018 to treat patients with a rare form of inherited blindness called biallelic RPE65 mutation-associated retinal dystrophy. The disease affects between 1,000 and 2,000 patients in the U.S. who have a mutation in both copies of a particular gene, RPE65. Luxturna delivers a normal copy of RPE65 to patients' retinal cells, allowing them to make a protein necessary for converting light to electrical signals and restoring their vision.

**STRIMVELIS:** About 15 patients are diagnosed in Europe every year with severe immunodeficiency from a rare inherited condition called adenosine deaminase deficiency (ADA-SCID). These patients' bodies cannot make the ADA enzyme, which is vital for healthy white blood cells. Strimvelis, approved in the E.U. in 2016, works by introducing the gene responsible for producing ADA into stem cells taken from the patient's own marrow. The cells are then reintroduced into the patient's bloodstream, where they are transported to the bone marrow and begin producing normal white blood cells that can produce ADA.

**YESCARTA:** Developed to treat a cancer called large B cell lymphoma, Yescarta was approved by the FDA in 2017 and in the E.U. in 2018. It is in clinical trials in China. Large B cell lymphoma affects white blood cells called lymphocytes. The treatment, part of an approach known as

CAR-T cell therapy, uses a virus to insert a gene that codes for proteins called chimeric antigen receptors (CARs) into a patient's T cells. When these cells are reintroduced into the patient's body, the CARs allow them to attach to and kill cancer cells in the bloodstream.

**ZOLGENSMA:** In May 2019 the FDA approved Zolgensma for children younger than two years with spinal muscular atrophy, a neuromuscular disorder that affects about one in 10,000 people worldwide. It is one of the leading genetic causes of infant mortality. Zolgensma delivers a healthy copy of the human SMN gene to a patient's motor neurons in a single treatment.

**ZYNTGLO:** Granted approval in the E.U. in May 2019, Zynteglo treats a blood disorder called beta thalassemia that reduces a patient's ability to produce hemoglobin, the protein in red blood cells that contains iron, leading to life-threatening anemia. The therapy has been approved for individuals 12 years and older who require regular blood transfusions. It employs a virus to introduce healthy copies of the gene for making hemoglobin into stem cells taken from the patient. The cells are then reintroduced into the bloodstream and transported to the bone marrow, where they begin producing healthy red blood cells that can manufacture hemoglobin.

## GENE INTERFERENCE

This approach uses a synthetic strand of RNA or DNA (called an oligonucleotide) that, when introduced into a patient's cell, can attach to a specific gene or its messenger molecules, effectively inactivating them. Some treatments use an antisense method, named for one DNA strand, and others rely on small interfering RNA strands, which stop instruction molecules that go from the gene to the cell's protein factories.

**DEFITELIO:** This drug contains a mixture of single-strand oligonucleotides obtained from the intestinal mucosa of pigs. It was approved (with limitations) in the U.S. and the E.U. in 2017 to treat severe cases of veno-occlusive disease, a disorder in which the small veins of the liver become obstructed, in patients who have received a bone marrow transplant.

**EXONDYS 51:** In 2016 the FDA granted approval to Exondys 51 amid some controversy regarding its efficacy; two members of the FDA review panel resigned in protest of the decision. The therapy is designed to treat a form of Duchenne muscular dystrophy caused by mutations in the RNA that codes for the protein that helps to connect muscle fibers' cytoskeletons to a surrounding matrix.

Exondys 51 is effective in treating about 13 percent of the Duchenne population.

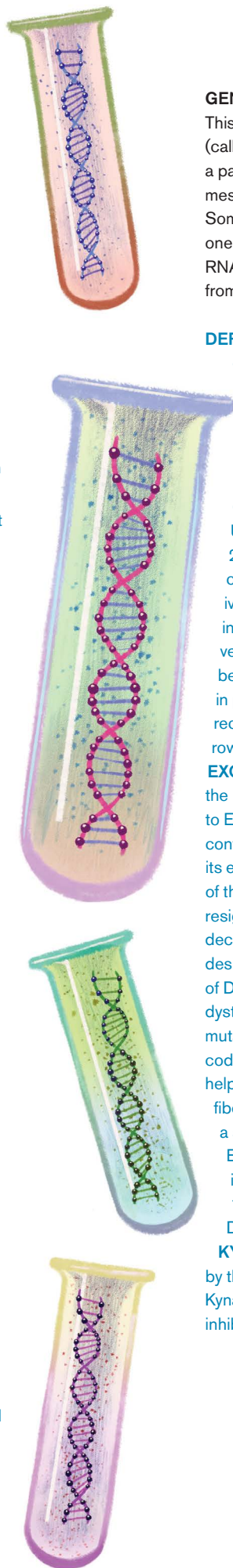
**KYNAMRO:** Approved by the FDA in 2013, Kynamro is designed to inhibit—or effectively shut

down production of—a protein that helps to produce low-density lipoprotein (LDL). Injected subcutaneously, this therapy is used to lower LDL levels in patients who have dangerously high cholesterol.

**MACUGEN:** Age-related macular degeneration is the leading cause of vision loss in people age 60 and older. It is caused by deterioration of the center of the retina due to leaking blood vessels. Approved in the U.S., Macugen inhibits these blood vessels from growing under the retina, thus treating the disorder.

**SPINRAZA:** With its FDA approval in 2016, Spinraza became the first gene-based therapy for spinal muscular atrophy. The inherited disorder is caused by low levels of SMN, a key protein for the maintenance of motor neurons. Spinraza binds to RNA from a "backup" gene called SMN2, converting that RNA into instructions for making fully functioning SMN proteins.

**Jim Daley** is a freelance journalist based in Chicago.



BIG DATA

## All of Us

DNA-based medicine needs more diversity to avoid harmful bias. One big research project is fixing that

By Stephanie Devaney

**WHEN THE RACE TO** sequence the first human genome was rushing toward the finish line about 20 years ago, I remember feeling mesmerized by what was about to happen. It was the dawn of a new century, and it seemed we were on the cusp of unlocking the meaning behind the blueprint of life, DNA. Once we could line up all 3.1 billion base pairs of the molecule in our genome, I thought—I was an undergraduate student at the time, dazzled by science—we would understand everything there is to know about human health and disease.

What I didn't know was that those first decades of genetic medicine would leave a lot of people behind. So I was taken aback several years later, in 2009, just after I got my doctorate in molecular genetics, when researchers at Duke University

reported that 96 percent of the genomic data we had gathered came from people of European ancestry. This was not the result of small numbers: they calculated the percentage using the more than 1.7 million individual genome samples analyzed at the time, but the samples were lacking diversity. Over the next few years things did not get much better, and as recently as four years ago genomic databases were still way out of balance, with more representation of Europeans and less of everyone else.

This inequity, if it is not fixed, will turn into tremendous health inequality. Today more and more people are getting answers about the underlying causes of their diseases because of medicine's ability to mine their genomes. There are hundreds of drugs that contain genetic information in their labeling because gene variants affect how bodies process these drugs, and knowing the variants that patients have helps doctors set the most beneficial dose for their patients. Moreover, today improved knowledge about the genomic drivers of different cancers has paid dividends in how physicians diagnose and treat many tumors. Yet people who are not white and not male have different sets of genes that do not always fit into these treatment regimens.

For example, African-Americans and Latinos have the highest rate of asthma in the U.S., but studies show that common drugs used in inhalers do not help them as well as they help whites. Asians who take the antiseizure drug carbamazepine have a higher risk of a severe, sometimes fatal, reaction. Nobody developing these drugs, or prescribing them when they first came into use, anticipated these problems. If DNA is one important factor in our quest for more effective medical treatment, we need to address the lack of diversity in genetic data.

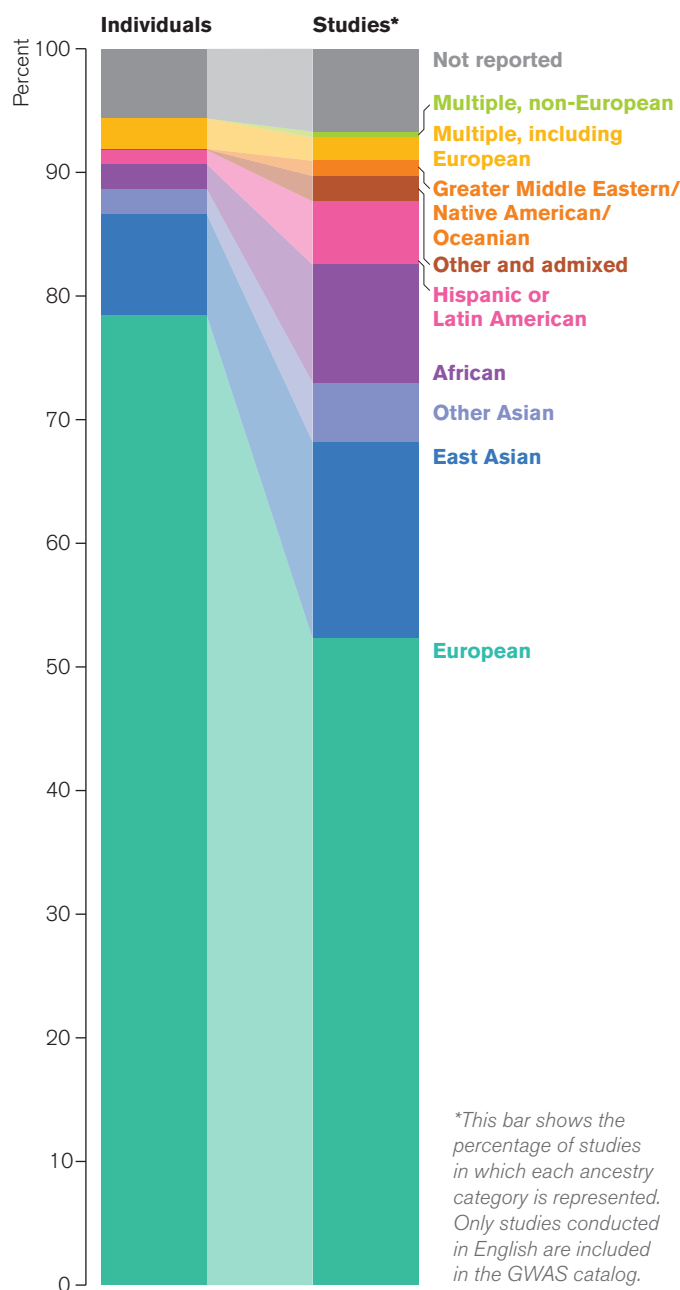
That is where the *All of Us* Research Program, where I work, hopes to help. Set up by the National Institutes of Health and launched in 2018, we are asking a million or more people from all backgrounds to join us as partners in research, not as human subjects, and share all kinds of health information over the course of their lives. Already we have more than 250,000 participants. More than 51 percent belong to racial and ethnic minorities, more than 10 percent are sexual and gender minorities, and overall more than 80 percent represent a group that has been historically underrepresented in research data sets.

People can join All of Us by going to our program Web site ([www.joinallofus.org](http://www.joinallofus.org)) and clicking "Join Now." After agreeing to participate, respondents can offer us their medical records, answer a variety of surveys about their health

## Biased Gene Studies

To link genes to disease risk and other traits, hundreds of genome-wide association studies (GWAS) have looked at the DNA of thousands of different people as of 2018. But in terms of racial background, these people are not so different. Taking all the projects together, 78 percent of the people in them are white Europeans, whereas just 2 percent are African and 1 percent are Hispanic or Latin American. The studies themselves also predominantly focused on Europeans and rarely on other populations. So gene variants that appear in non-European people and may be linked to illness rarely show up in this research. The scarcity makes it hard to analyze and understand the significance of the variants.

**Racial Backgrounds in Published Gene-Association Studies**



and lifestyle, and participate in other activities such as syncing their fitness tracker data to our program. We also have hundreds of enrollment sites at local hospitals and health centers across the country where participants can provide samples of blood and urine to help researchers study their DNA. Our hope is for people to stick with us for 10 years or more because, as the program grows, we will regularly add new ways for them to learn about themselves and contribute to research.

### THE MOMENT IS RIGHT

**A LOT OF THIS PARTICIPANT-RESEARCHER** collaboration is linked to advances in technology. Sequencing that first human genome had a \$1-billion price tag. Today such a sequence costs less than \$1,000 and can take less than 24 hours to complete. It is also easier to integrate this information with other crucial medical data. Health care organizations have been turning their patients' paper-based medical records into electronic versions. As of 2017, 96 percent of all U.S. hospitals and 80 percent of all office-based doctors are using a certified electronic health record system. New apps on smartphones and other digital health technologies such as smart watches collect data from nearly anywhere and directly from a person. These trends all make it easier to store, share and mine large data sets for answers to questions about disease causes and effects. Such trends also raise big and disturbing issues about privacy, making it important for projects such as ours to have both strong security and full transparency to all our participants.

And it is crucial to treat these people as partners. The actions of past medical researchers have earned much distrust in minority communities, after causing harm in the Tuskegee Syphilis Study, where researchers misled African-American men with syphilis and never gave them adequate treatment, and with the widespread use of HeLa cells, which were taken from a patient named Henrietta Lacks without her knowledge or permission. People wanted to see research go forward but *with* them rather than about them. To overcome this kind of distrust, All of Us is using a new model for research, one that invites input from participants as well as researchers with science degrees. Participants serve on the program's advisory and governing bodies, working groups, and task forces. We have also partnered with local health care organizations, hospitals, and community groups to advise us and help find people to participate. Community engagement is not familiar ground for large medical research projects, and we are still learning the best ways to do it.

Some studies have provided us with blueprints for developing long-term relationships like the ones we hope to have, studies that have changed medicine for the better. The Framingham Heart Study, for example, started in 1948 with 5,209 men and women, largely white, from one town in Massachusetts. With a 99 percent retention rate, the study continues to this day. As participants share data year after

year, researchers can see how their heart health changes over time. The risk factors for heart disease identified by the Framingham study—such as high blood pressure, high cholesterol, smoking and obesity—are so ingrained in our collective consciousness and our approach to health care that they feel like common sense.

### GOING FURTHER

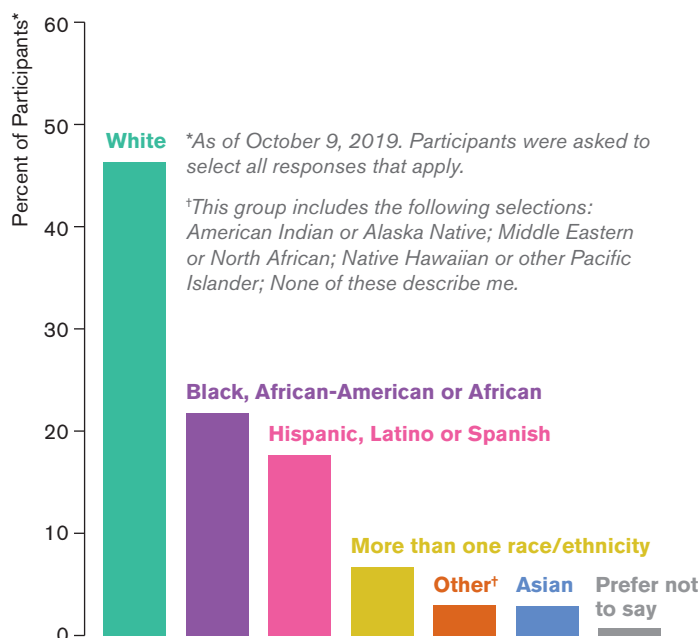
**THIS KIND OF MEDICAL DISCOVERY** is what we envision for All of Us, but we want to take it further, with participants who are not all white and who represent diversity in many dimensions, not just traditional race labels that, in reality, encompass a lot of different backgrounds. If we're going to get at the root causes of health and disease, this means understanding the differences and similarities among us all. For example, sickle cell disease occurs when someone inherits two mutated genes for the oxygen-carrying protein hemoglobin. It affects 100,000 African-Americans and more than 20 million people around the world. In contrast, sickle cell trait—meaning just one of these genes is mutated—actually gives people an advantage in surviving malaria, which makes evolutionary sense if your ancestors came from areas such as Africa where malaria is prevalent. New studies, however, have found that sickle cell trait might not be as benign as doctors used to believe, because it may increase the risk for kidney disease. Some African-Americans are more susceptible to this risk and some less. There's clearly more to learn about why this might be the case and about how different DNA variants might interact to affect the health of people with sickle cell trait. The DNA information from more than a million All of Us participants could help researchers learn much more about complex traits like this.

We do have to start with some of the broad-brush categories to recruit enough people to start recognizing the more fine-grained groups among them. Currently we are exceeding our goal of overrepresenting groups that have been historically underrepresented in research. For instance, African-Americans make up about 13 percent of the U.S. population but just 3 percent of the samples previously used in genome studies. In All of Us, 21.5 percent of participants so far are African-American. Similarly, Hispanics constitute about 18 percent of the U.S. population but in 2016 made up less than 1 percent of the data in our genomic databases. Today 17.6 percent of All of Us participants are Hispanic.

That diversity will help us discover more about how DNA affects health across different communities, but the molecule will not be our sole focus. Many factors beyond our genes are at play when it comes to disease. We know that where you were born, what you eat, the stress you feel, and other clinical and biological factors affect health, but we still don't understand by how much. For example, when we think about some of the most common chronic diseases that afflict our population—high blood pressure is one ex-

## A Better Balance

A new precision medicine project, All of Us, has much larger populations of groups that have been historically underrepresented in genetics research. The project, sponsored by the U.S. National Institutes of Health, began recruiting participants in 2018. More than 250,000 people enrolled by October 2019, and just over 20 percent are black, African-American or African. About 18 percent are Latino, Hispanic or Spanish. Nearly 3 percent are Asian, and 6.7 percent are of mixed races. Slightly less than half of the people are white. The project's goal is to get DNA and other health information from more than one million people.



ample—many of them disproportionately affect the most socially and economically disadvantaged people in our country. And from what we can tell at the moment, the determinants are not simply their race or ethnicity. Risks also include family structure, socioeconomic status, stressors such as trauma, sex and gender inequality, availability of nutrient-rich foods, access to health care, and many other factors that we can capture in the All of Us data set.

Within the next several years, we should be able to compare this rich set of information with participants' DNA. When we do so, scientists such as myself, the All of Us participants and all of you will start to get a clearer picture of the roles that biology and environment play in disease development, and—most important of all—what we can do about it.

Molecular geneticist **Stephanie Devaney** is deputy director of the All of Us research program at the National Institutes of Health. She was the staff lead for the White House Precision Medicine initiative.